ADVANCES IN INTELLIGENT SYSTEMS AND COMPUTING 190

Rudolf Kruse
Michael R. Berthold
Christian Moewes
María Ángeles Gil
Przemysław Grzegorzewski
Olgierd Hryniewicz (Eds.)

Synergies of Soft Computing and Statistics for Intelligent Data Analysis

Springer

cajAstur

Advances in Intelligent Systems and Computing 190

Editor-in-Chief

Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

For further volumes:
http://www.springer.com/series/11156

Rudolf Kruse, Michael R. Berthold,
Christian Moewes, María Ángeles Gil,
Przemysław Grzegorzewski,
and Olgierd Hryniewicz (Eds.)

Synergies of Soft Computing and Statistics for Intelligent Data Analysis

Springer

Editors
Prof. Dr. Rudolf Kruse
Otto-von-Guericke University of Magdeburg
Faculty of Computer Science
Magdeburg
Germany

Prof. Dr. María Ángeles Gil
Department of Statistics and OR
University of Oviedo
Oviedo
Spain

Prof. Dr. Michael R. Berthold
University of Konstanz
FB Informatik & Informationswissenschaft
Konstanz
Germany

Prof. Dr. Przemysław Grzegorzewski
Polish Academy of Sciences
Systems Research Institute
Warsaw
Poland

Dipl.-Inform. Christian Moewes
Otto-von-Guericke University of Magdeburg
Faculty of Computer Science
Magdeburg
Germany

Prof. Dr. Olgierd Hryniewicz
Polish Academy of Sciences
Systems Research Institute
Warsaw
Poland

ISSN 2194-5357
ISBN 978-3-642-33041-4
DOI 10.1007/978-3-642-33042-1
Springer Heidelberg New York Dordrecht London

e-ISSN 2194-5365
e-ISBN 978-3-642-33042-1

Library of Congress Control Number: 2012945296

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

We are proud to present the proceedings of the 6th International Conference on Soft Methods in Probability and Statistics. The conference took place in Konstanz, Germany, October 4–6, 2012. The SMPS conferences series started in 2002 in Warsaw and moved from Oviedo (2004), Bristol (2006), Toulouse (2008) to Mieres-Oviedo (2010). SMPS'2012 was organized by the Computational Intelligence group in Magdeburg and the Intelligent Data Analysis group in Konstanz. The theme of SMPS2012 was "Synergies of Soft Computing and Statistics for Intelligent Data Analysis".

The main objective of the SMPS conference series is to strengthen the dialogue between various research communities in the field of data analysis in order to cross-fertilize the fields and generate mutually beneficial activities. In SMPS'2012 we were especially interested in bringing experts from the areas of Soft Computing and Statistics together. Both branches have different intentions on data analysis as they stem from computer science and mathematics, respectively. Soft computing is able to quickly produce low-cost solutions using nature-inspired problem-solving strategies. Its ability to adapt to different problems and models led to its success in real-world applications. Also, its inherent necessity to construct understandable and interpretable solutions made soft computing very popular in economical fields. The field of statistics aims at much less subjective goals. It focuses on the need for mathematical methods that validate models and ensure their applicability based on observations maximizing some kind of likelihood. It is our hope that the synergies of both fields improve intelligent data analysis methods in terms of robustness to noise and applicability to larger datasets, while being able to efficiently obtain understandable solutions of real-world problems.

SMPS'2012 provided an attractive interdisciplinary forum for discussions and mutual exchange of knowledge in the field of intelligent data analysis. The 58 papers in this volume were carefully selected by an extensive reviewing process. Every paper has been reviewed by three of 113 international experts. We are delighted that Christian Borgelt, principal researcher at the European Centre for Soft Computing, Lawrence O'Higgins Hall, professor of

computer science at the University of South Florida, and Hannu T. Toivonen, professor of computer science at the University of Helsinki, accepted our invitation to present keynote lectures. Part I of the volume contains these invited papers. Part II encloses contributions to the foundations of uncertainty theories including imprecise probability, probability theory and fuzzy set theory. Part III consists of a variety of papers dealing with soft statistical methods ranging from statistical inference to statistical tests. Part IV focuses on mathematical aspects of soft methods applied to probability and statistics, e.g. copulas, decision making, partial knowledge and conditional uncertainty measures. And Part V finally comprises application-orientated papers devoted to engineering. The methods described here exploit information mining, machine learning techniques and computational intelligence. Most applications stem from bioinformatics, human sciences and automobile industry.

The editors are very grateful to all contributing authors, invited speakers, program committee members, and additional referees who made it possible to put together an attractive program for the conference. This conference has benefited from the financial support of the Spanish bank CajAstur, which covered all the production and distribution costs of the proceedings. We thank the editor of the Springer series Advances in Soft Computing, Janusz Kacprzyk, and Springer-Verlag for the dedication to the production of this volume. We are particularly grateful to the universities of Konstanz and Magdeburg as well as the German Society of Computer Science and the European Society for Fuzzy Logic and Technology for their continuous support. Finally, we thank Heather Fyson who did an outstanding job putting the conference on the ground in Konstanz. Months before this conference, she already started working on everything that made this conference run smoothly and so enjoyable for all attendees.

Konstanz and Magdeburg
July 2012

Rudolf Kruse
Michael R. Berthold
Christian Moewes
María Ángeles Gil
Przemysław Grzegorzewski
Olgierd Hryniewicz

Organization

General Chairs

Michael R. Berthold (University of Konstanz, Germany)
Rudolf Kruse (University of Magdeburg, Germany)

Advisory Committee (Core SMPS Group)

María Ángeles Gil (University of Oviedo, Spain)
Przemysław Grzegorzewski (Polish Academy of Sciences, Warsaw, Poland)
Olgierd Hryniewicz (Polish Academy of Sciences, Warsaw, Poland)

Program Committee

Bernard de Baets (Ghent University, Belgium)
Michael Beer (University of Liverpool, United Kindom)
Radim Belohlavek (Palacký University, Olomouc, Czech Republic)
Piero Bonissone (GE Global Research, Niskayuna, NY, USA)
Christian Borgelt (European Centre for Soft Computing, Mieres, Spain)
Bernadette Bouchon-Meunier (Université Pierre et Marie Curie, Paris, France)
Giulianella Coletti (University of Perugia, Italy)
Ana Colubi (University of Oviedo, Spain)
Gert de Cooman (Ghent University, Belgium)
Ines Couso (University of Oviedo, Spain)
Thierry Denœux (Université de Technologie de Compiègne, France)
Sébastien Destercke (Université de Technologie de Compiègne, France)
Didier Dubois (Université Paul Sabatier, Toulouse, France)
Fabrizio Durante (Free University of Bozen, Bolzano, Italy)
Peter Filzmoser (TU Vienna, Austria)
Jörg Gebhardt (Volkswagen AG, Wolfsburg, Germany)
Lluís Godo Lacasa (Universitat Autònoma de Barcelona, Spain)
Gil González-Rodríguez (University of Oviedo, Spain)

Michel Grabisch (Université Panthéon-Sorbonne, France)
Mario Guarracino (Second University of Naples, Italy)
Frank Hoffmann (TU Dortmund, Germany)
Eyke Hüllermeier (University of Marburg, Germany)
Janusz Kacprzyk (Polish Academy of Sciences, Warsaw, Poland)
Uzay Kaymak (Eindhoven University of Technology, The Netherlands)
Etienne Kerre (Ghent University, Belgium)
Vladik Kreinovich (University of Texas at El Paso, TX, USA)
Jonathan Lawry (University of Bristol, United Kingdom)
Marie-Jeanne Lesot (Université Pierre et Marie Curie, Paris, France)
Shoumei Li (Beijing University Of Technology, China)
Edwin Lughofer (Johannes Kepler University, Linz, Austria)
Ralf Mikut (Karlsruhe Institute of Technology, Germany)
Serafin Moral (University of Granada, Spain)
Detlef D. Nauck (British Telecom, Ipswich, United Kindom)
Mirko Navara (Czech Technical University, Praha, Czech Republic)
Vilem Novak (University of Ostrava, Czech Republic)
Henri Prade (Université Paul Sabatier, Toulouse, France)
Dan Ralescu (University of Cincinnati, USA)
Thomas A. Runkler (Siemens AG, Munich, Germany)
Enrique Ruspini (European Centre for Soft Computing, Mieres, Spain)
João Miguel da Costa Sousa (Technical University of Lisbon, Portugal)
Wolfgang Trutschnig (European Centre for Soft Computing, Mieres, Spain)
Bülent Tütmez (İnönü University, Malatya, Turkey)
Stefan Van Aelst (Ghent University, Belgium)
Michel Verleysen (Université catholique de Lovain, Louvain-la-Neuve, Belgium)
Peter Winker (Justus-Liebig-Universität Gießen, Germany)
Marco Zaffalon (University of Lugano, Switzerland)

Additional Referees

Andreas Alfons, Andrés David Báez Sánchez, Pablo Bermejo, Holger Billhardt, Angela Blanco-Fernández, Christian Braune, Andrea Capotorti, María Rosa Casals Varela, Andrea Cerioli, Yun-Shiow Chen, Norberto Corral, Pierpaolo D'Urso, Serena Doria, Luca Fanelli, Juan Fernández Sánchez, Maria Brigida Ferraro, Rachele Foschi, Enrico Foscolo, Marta García-Bárzana, Luis Angel García Escudero, Roberto Ghiselli Ricci, Angelo Gilio, Paolo Giordani, Pascal Held, Mehran Hojati, Karel Hron, Mia Hubert, Jui-Chung Hung, Piotr Jaworski, Huidong Jin, Radim Jirousek, Mehdi Khashei, Gernot D. Kleiter, Tomáš Kroupa, Alexander Lepskiy, Baoding Liu, María Teresa Lopez García, Kevin Loquin, María Asunción Lubiano, Marek T. Malinowski, Peter McCullagh, Deyu Meng, Radko Mesiar, Mariusz Michta, Enrique Miranda, Stephan Morgenthaler, Zoran Ognjanović, Roberta Pappadà, Franco Pellerey, Domenico Perrotta, Georg Peters, Davide Petturiti, Niki Pfeifer,

Florin Popescu, Erik Quaeghebeur, Ana Belén Ramos-Guajardo, Giuliana Regoli, Marco Riani, José Antonio Rodriguez Lallena, Peter Sarkoci, Carlo Sempi, Fabio Spizzichino, Olivier Strauss, Pedro Terán, Valentin Todorov, Matthias Troffaes, Lev Utkin, Alain C. Vandal, Barbara Vantaggi, Jinping Zhang

Local Organization

Heather Fyson (University of Konstanz, Germany)

Publication Chair

Christian Moewes (University of Magdeburg, Germany)

Special Sessions Chair

Frank Klawonn (Ostfalia University of Applied Sciences, Wolfenbüttel, Germany)

Publicity Chair

Andreas Nürnberger (University of Magdeburg, Germany)

Contents

Probability and Fuzzy Sets

Part III: Statistical Methods

Statistics with Imprecise Data

Fuzzy Random Variables

Fuzzy Regression Methods

Soft Computing and Statistics

Robust Statistics

Computational Intelligence

Industrial Applications

Machine Learning

Part I
Invited Papers

Soft Pattern Mining in Neuroscience

Christian Borgelt

Abstract. While the lower-level mechanisms of neural information processing (in biological neural networks) are fairly well understood, the principles of higher-level processing remain a topic of intense debate in the neuroscience community. With many theories competing to explain how stimuli are encoded in nerve signal (spike) patterns, data analysis tools are desired by which proper tests can be carried out on recorded parallel spike trains. This paper surveys how pattern mining methods, especially soft methods that tackle the core problems of *temporal imprecision* and *selective participation*, can help to test the *temporal coincidence coding hypothesis*. Future challenges consist in extending these methods, in particular to the case of *spatio-temporal coding*.

1 Introduction

Basically all information transmission and processing in humans and animals is carried out by the *nervous system*, which is a network of special cells called *neurons* or *nerve cells*. These cells communicate with each other by electrical and chemical signals. While the lower-level mechanisms are fairly well understood (see Section 2) and it is widely accepted in the neuroscience community that stimuli are encoded and processed by cell assemblies rather than single cells [17, 23], it is still a topic of intense ongoing debate how exactly information is encoded and processed on such a higher level: there are many competing theories, each of which has its domain of validity. Due to modern multi-electrode arrays, which allow to record the electrical signals emitted by hundreds of neurons in parallel [8], more and more data becomes available in the form of (massively) *parallel spike trains* that can help to tackle the challenge of understanding higher-level neural information processing.

Christian Borgelt
European Centre for Soft Computing, Edificio de Investigación, 33600 Mieres, Asturias, Spain
e-mail: christian@borgelt.net

R. Kruse et al. (Eds.): Synergies of Soft Computing and Statistics, AISC 190, pp. 3–10.
springerlink.com

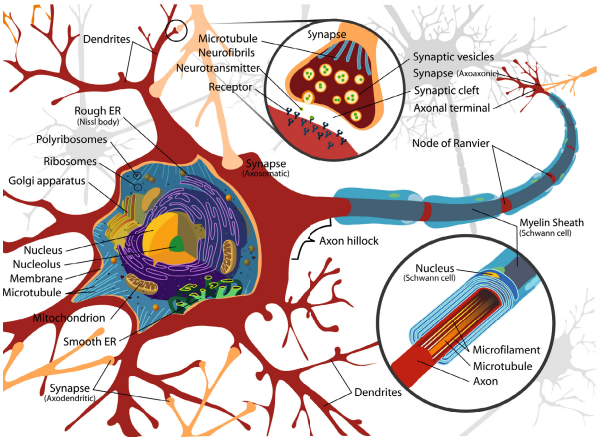

Fig. 1 Diagram of a typical myelinated vertebrate motoneuron (source: Wikipedia [27]), showing the main parts involved in its signaling activity like the *dendrites*, the *axon*, and the *synapses*

After reviewing some of the main competing models of neural information coding (Section 2), this paper focuses on the *temporal coincidence coding hypothesis*. It explores how pattern mining methods can help in the search for synchronous spike patterns in parallel spike trains (Section 3) and considers, in particular, soft methods that can handle the core problems of *temporal imprecision* and *selective participation* (Section 4). The paper closes with an outlook on future work, especially tackling the challenge of identifying *spatio-temporal patterns* under such conditions (Section 5).

2 Neural Information Processing

Essentially, neurons are electrically excitable cells that send signals to each other. The mechanisms are well understood on a physiological and chemical level, but how several neurons coordinate their activity is not quite clear yet.

Physiology and Signaling Activity. Neurons are special types of cells that can be found in most animals. They connect to each other, thus forming complex networks. Attached to the *cell body* (or *soma*) are several arborescent branches that are called *dendrites* and one longer cellular extension called the *axon*. The axon terminals form junctions, so-called *synapses*, with the dendrites or the cell bodies of other neurons (see Figure 1) [14].

The most typical form of communication between neurons (this is a *very* simplified description!) is that the axon terminals of a neuron release chemical substances, called *neurotransmitters*, which act on the membrane of the connected neuron and change its polarization (its electrical potential). Synapses that reduce the potential difference between the inside and the outside of the membrane are called *excitatory*, those that increase it, *inhibitory*. Although the change caused by a single synapse is comparatively small, the effects of

multiple synapses accumulate. If the total excitatory input is large enough, the start of the axon becomes, for a short period of time (around 1ms), depolarized (i.e. the potential difference is inverted). This sudden change of the electrical potential, called *action potential*, travels along the axon, with the speed depending on the amount of *myelin* present. When this nerve impulse reaches the end of the axon, it triggers the release of neurotransmitters. Thus the signal is passed on to the next neuron [14]. The electrical signals can be recorded with electrodes, yielding so-called *spike trains*.

Neural Information Coding. It is widely accepted in the neuroscience community that stimuli and other pieces of information are not represented by individual neurons and their action potentials, but that multiple neurons work together, forming so-called *cell assemblies*. However, there are several competing theories about how exactly the information is encoded. The main models that are considered include, but are not limited to the following [23]:

- **Frequency Coding** [29, 12]
  Neurons generate spikes trains with varying frequency as a response to different stimulus intensities: the stronger the stimulus, the higher the spike frequency. Frequency coding is used in the motor system, which directly or indirectly controls muscles, because the rate at which a muscle contracts is correlated with the number of spikes it receives. Frequency coding has also been shown to be present in the sensory system.
- **Temporal Coincidence Coding** [21, 30, 19, 25]
  Tighter coincidence of spikes recorded from different neurons represent higher stimulus intensity, with spike occurrences being modulated by local field oscillation [23]. A temporal coincidence code has the advantage that it leads to shorter "switching times," because it avoids the need to measure a frequency, which requires to observe multiple spikes. Therefore it appears to be a better model for neural processing in the cerebral cortex.
- **Delay Coding** [18, 9]
  The input stimulus is converted into a spike delay (possibly relative to some reference signal). A neuron that is stimulated more strongly reaches the depolarization threshold earlier and thus initiates a spike (action potential) sooner than neurons that are stimulated less strongly.
- **Spatio-Temporal Coding** [2, 1]
  Neurons emit a causal sequence of spikes in response to a stimulus configuration. A stronger stimulus induces spikes earlier and initiates spikes in other, connected cells. The sequence of spike propagation is determined by the spatio-temporal configuration of the stimulus as well as the connectivity of the network [23]. This coding model can be seen as integrating the temporal coincidence and the delay coding principles.

Among other models a spatio-temporal scheme based on a frequency code [28] is noteworthy. In this model the increased spike frequencies form specific spatio-temporal patterns over the involved neurons. Thus it can be seen as combining spatio-temporal coding with frequency coding.

3 Detecting Synchronous Activity

This paper focuses on the temporal coincidence coding hypothesis and thus on the task to detect unusual synchronous spiking activity in recorded parallel spike trains, where "unusual" means that it cannot be explained as a chance event. In addition, we do not merely consider whether a parallel spike train contains synchronous spiking activity (e.g. [31]) or whether a given neuron participates in the synchronous spiking activity of a cell assembly (of otherwise unknown composition) (e.g. [4]). Rather we concentrate on the most complex task of identifying specific assemblies that exhibit(s) (significant) synchronous spiking activity (e.g. [13, 3]). Tackling this task is computationally expensive for (massively) parallel spike trains due to a combinatorial explosion of possible neuron groups that have to be examined.

Other core problems are *temporal imprecision* and *selective participation*. The former means that it cannot be expected that spikes are temporally perfectly aligned, while the latter means that only a subset of the neurons in an assembly may participate in any given synchronous spiking event, with the subset varying between different such events. Note that both may be the effect of deficiencies of the spike recording process (the spike time or even whether a spike occurred is not correctly extracted from the measured profile of the electrical potential) or may be due to the underlying biological process (delays or even failures to produce a spike due to lower total synaptic input, as neurons may receive signals coding different information in parallel).

The most common (or even: the almost exclusively applied) method of handling temporal imprecision is time binning: given a user-specified bin width, a spike train, which is originally a (continuous) point process of spike times, is turned into a binary sequence: a 1 means that the corresponding neuron produced a spike and a 0 that there is no spike in the corresponding time bin. In this way the problem is essentially transformed into a frequent item set mining problem [3]. The translation of the relevant notions to market basket analysis (for which frequent item set mining was originally developed) and to spike train analysis is shown in Table 1. Clearly, the problems are structurally equivalent and thus can be attacked with the same means.

The standard problem of frequent item set mining—namely that a huge number of frequent item sets may be found, most of them *false discoveries*—is best addressed by randomization methods [22, 15]. In spike train analysis, these methods take the form of surrogate data generation schemes, since one tries to preserve as many properties (that are deemed biologically relevant, e.g. inter-spike intervals) as possible, while destroying the coincidences. A survey of such surrogate data generation methods can be found in [20].

In essence, an assembly detection method then works as follows: a sufficient number of surrogate data sets (say, 1000 or 10,000) are created and mined for frequent item sets, which are identified by their size (number of neurons) and support (number of coincidences). Then the original data set is mined and if patterns of a size and support (but ignoring the exact composition by

Table 1 Translation of basic notions of frequent item set mining to market basket analysis (for which it was originally developed) and to spike train analysis

mathematical problem	market basket analysis	spike train analysis
item	product	neuron
item base	set of all products	set of all neurons
— (transaction id)	customer	time bin
transaction	set of products bought by a customer	set of neurons firing in a time bin
frequent item set	set of products frequently bought together	set of neurons frequently firing together

neurons) can be found that do not show up in any of the surrogate data sets, these patterns can be considered significant results.

4 Soft Pattern Mining

Accepting time binning for now as a simple (though deficient, see below) method for handling temporal imprecision, let us turn to the problem of selective participation. In the framework of frequent item set mining this is a well-known problem for which many approaches exist (see, e.g., [5]). The core idea is this: in standard frequent item set mining a transaction (time bin) supports an item set (neuron set) only if all items in the set are present. By relaxing the support definition, allowing for some items of a given set to be missing from a transaction, we arrive at *fault-tolerant item set mining*. The various algorithms for this task can be roughly categorized into (1) error-based approaches, which allow for a maximum number of missing items, (2) density-based approaches, which allow for a maximum fraction of missing items, and (3) cost-based approaches, which reduce the support contribution of a transaction depending on the number of missing items (and may, in addition, restrict the number of missing items) [5].

However, such approaches suffer from the even larger search space (as more item sets need to be examined) and thus can increase the computational costs considerably. An alternative approach that avoids an exhaustive enumeration relies on distance measures for binary vectors [10] and uses multi-dimensional scaling [11] to a single dimension to group neurons together that exhibit similar spiking activity [7]. The actual assemblies are then discovered by traversing the neurons according to their image location and testing for dependence. The approach of computing distances of time-binned spike trains has been extended to various well-known clustering methods in [6].

All of the mentioned methods work on time binned data. However, the time binning approach has several severe drawbacks. In the first place, the induced concept of synchrony is two-valued, that is, spikes are either synchronous

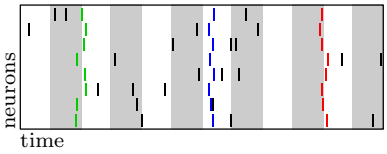

Fig. 2 Eight parallel spike trains with three coincident spiking events (shown in color), two of which are disrupted by time bin boundaries (time bins indicated by gray and white stripes)

(namely if they lie in the same time bin) or not. We have no means to express that the spikes of some coincident event are better aligned than those of another. Secondly, time binning leads to anomalies: two spikes that are (very) close together in time, but happen to be on different sides of a time bin boundary are seen as not synchronous, while two spikes that are almost as far apart as the length of a time bin, but happen to fall into the same time bin, are seen as synchronous. Generally, the location of the time bin boundaries can have a disruptive effect. This is illustrated in Figure 2, where two of the three coincidences of the eight neurons (shown in color) cannot be detected, because they are split by badly placed time bin boundaries.

These problems have been addressed with the influence map approach (see [24, 6]), which bears some resemblance to the definition of a distance measure for continuous spike trains suggested in [26]. The core idea is to surround each spike time with an influence region, which specifies how imprecisely another spike may be placed, which is still to be considered as synchronous. Thus one can define a graded notion of synchrony based on the (relative) overlap of such influence regions. Unfortunately, a direct generalization of binary distance measures to this case (using properly scaled durations instead of time bin counts) seems to lose too much information due to the fact that full synchrony can only be achieved with perfectly aligned spikes [6].

As a solution one may consider specific groups of spikes, one from each neuron, rather than intersecting, over a set of neurons, the union of the influence regions of the spikes of each neuron. This allows to define ϵ-tolerant synchrony, which is 1 as long as the temporal imprecision is less than a user-specified ϵ and becomes graded only beyond that. In addition, extensions to the fault-tolerant case are possible by allowing some spikes to be missing.

5 Future Challenges

The methods reviewed in this paper were devised to detect synchronous activity. However, attention in the neuroscience community shifts increasingly towards spatio-temporal spike patterns as the more general concept, which contains synchronous spiking as a special case. If the time binning approach is accepted, frequent pattern mining offers readily available solutions, for example, in the form of the Spade [33] and cSpade algorithms [32]. However, these approaches require discretized time. Similarly, approaches developed in

the neuroscience community (e.g. [2]) are based on time bins, and thus suffer from the mentioned anomalies. In addition, these methods cannot handle faults, in the sense of individual missing spikes: they only count full occurrences of the potential patterns. It is a challenging, but very fruitful problem to extend these approaches (possibly with influence maps) to continuous time or find alternative methods that can handle both faults and continuous time.

Acknowledgements. I am grateful to Denise Berger, Christian Braune, Iván Castro-León, George Gerstein, Sonja Grün, Sebastien Louis, and David Picado-Muiño (listed alphabetically by last name, no precedence expressed), with whom I have cooperated on several of the topics described in this paper.

References

1. Abeles, M., Bergman, H., Margalit, E., Vaadia, E.: Spatiotemporal firing patterns in the frontal cortex of behaving monkeys. J. Neurophysiol. 70(4), 1629–1638 (1993)
2. Abeles, M., Gerstein, G.L.: Detecting spatiotemporal firing patterns among simultaneously recorded single neurons. J. Neurophysiol. 60(3), 909–924 (1988)
3. Berger, D., Borgelt, C., Diesmann, M., Gerstein, G., Grün, S.: An accretion based data mining algorithm for identification of sets of correlated neurons. BMC Neurosci. 10, 1–2 (2009)
4. Berger, D., Borgelt, C., Louis, S., Morrison, A., Grün, S.: Efficient identification of assembly neurons within massively parallel spike trains. Comput. Intell. Neurosci. 2010, 1–18 (2010)
5. Borgelt, C., Braune, C., Kötter, T., Grün, S.: New algorithms for finding approximate frequent item sets. Soft Comput 16(5), 903–917 (2012)
6. Braune, C.: Analysis of parallel spike trains with clustering methods. Master's thesis, Faculty of Computer Science, Otto-von-Guericke University of Magdeburg, Germany (2012)
7. Braune, C., Borgelt, C., Grün, S.: Finding Ensembles of Neurons in Spike Trains by Non-linear Mapping and Statistical Testing. In: Gama, J., Bradley, E., Hollmén, J. (eds.) IDA 2011. LNCS, vol. 7014, pp. 55–66. Springer, Heidelberg (2011)
8. Buzsáki, G.: Large-scale recording of neuronal ensembles. Nat. Neurosci. 7(5), 446–451 (2004)
9. Buzsáki, G., Chrobak, J.J.: Temporal structure in spatially organized neuronal ensembles: a role for interneuronal networks. Curr. Opin. Neurobiol. 5(4), 504–510 (1995)
10. Choi, S., Cha, S., Tappert, C.C.: A survey of binary similarity and distance measures. J. Systemics Cybern. Inform. 8(1), 43–48 (2010)
11. Cox, T.F., Cox, M.A.A.: Multidimensional Scaling, 2nd edn. Chapman & Hall, London (2001)
12. Eccles, J.C.: The Physiology of Nerve Cells. Johns Hopkins University Press, Baltimore (1957)
13. Gerstein, G.L., Perkel, D.H., Subramanian, K.N.: Identification of functionally related neural assemblies. Brain Res. 140(1), 43–62 (1978)
14. Gilman, S., Newman, S.: Essentials of Clinical Neuroanatomy and Neurophysiology, 10th edn. F.A. Davis Company, Philadelphia (2002)

15. Gionis, A., Mannila, H., Mielikäinen, T., Tsaparas, P.: Assessing data mining results via swap randomization. ACM Trans. Knowl. Discov. Data 1(3) (2007)
16. Grün, S., Rotter, S. (eds.): Analysis of Parallel Spike Trains. Springer Series in Computational Neuroscience, vol. 7. Springer US, New York (2010)
17. Hebb, D.O.: The Organization of Behavior: A Neuropsychological Theory. John Wiley & Sons, Ltd., New York (1949)
18. Hopfield, J.J.: Pattern recognition computation using action potential timing for stimulus representation. Nature 376(6535), 33–36 (1995)
19. König, P., Engel, A.K., Singer, W.: Integrator or coincidence detector? the role of the cortical neuron revisited. Trends Neurosci. 19(4), 130–137 (1996)
20. Louis, S., Borgelt, C., Grün, S.: Generation and selection of surrogate methods for correlation analysis. In: Grün and Rotter [16], pp. 359–382
21. von der Malsburg, C., Bienenstock, E.: A neural network for the retrieval of superimposed connection patterns. Europhys. Lett. 3(11), 1243–1249 (1987)
22. Megiddo, N., Srikant, R.: Discovering predictive association rules. In: Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining (KDD 1998), New York, NY, pp. 274–278 (1998)
23. Nádasy, Z.: Spatio-temporal patterns in the extracellular recording of hippocampal pyramidal cells: From single spikes to spike sequences. Ph.D. thesis, Rutgers University, Newark, NJ, USA (1998)
24. Picado-Muiño, D., Castro-León, I., Borgelt, C.: Continuous-time characterization of spike synchrony and joint spiking activity in parallel spike trains (submitted, 2012)
25. Riehle, A., Grün, S., Diesmann, M., Aertsen, A.: Spike synchronization and rate modulation differentially involved in motor cortical function. Science 278(5345), 1950–1953 (1997)
26. van Rossum, M.C.: A novel spike distance. Neural Comput. 13(4), 751–763 (2001)
27. Ruiz Villarreal, M.: Complete neuron cell diagram (2007), http://commons.wikimedia.org/wiki/File:Complete_neuron_cell_diagram_en.svg
28. Seidemann, E., Meilijson, I., Abeles, M., Bergman, H., Vaadia, E.: Simultaneously recorded single units in the frontal cortex go through sequences of discrete and stable states in monkeys performing a delayed localization task. J. Neurosci. 16(2), 752–768 (1996)
29. Sherrington, C.S.: The Integrative Action of the Nervous System. Yale University Press, New Haven (1906)
30. Singer, W.: Synchronization of cortical activity and its putative role in information processing and learning. Annu. Rev. Physiol. 55, 349–374 (1993) PMID: 8466179
31. Staude, B., Rotter, S., Grün, S.: CuBIC: cumulant based inference of higher-order correlations in massively parallel spike trains. J. Comput. Neurosci. 29(1-2), 327–350 (2010)
32. Zaki, M.J.: Sequence mining in categorical domains: incorporating constraints. In: Proc. 9th Int. Conf. on Information and Knowledge Management, CIKM 2000, pp. 422–429. ACM, New York (2000)
33. Zaki, M.J.: SPADE: an efficient algorithm for mining frequent sequences. Mach. Learn. 42(1-2), 31–60 (2001)

Exploring Big Data with Scalable Soft Clustering

Lawrence O. Hall

Abstract. Sky surveys for Astronomy are expected to generate 2.5 petabytes a year. Electronic medical records hold the promise of treatment comparisons, grouping patients by outcomes but will be contained in petabyte data storage. We can store lots of data and much of it wont have labels. How can we analyze or explore the data? Unsupervised clustering, fuzzy, possibilistic or probabilistic will allow us to group data. However, the algorithms scale poorly in terms of computation time as the data gets large and are impractical without modification when the data exceeds the size of memory. We will explore distributed clustering, stream data clustering and subsampling approaches to enable scalable clustering. Examples will show that one can scale to build good models of the data without necessarily seeing all the data and, if needed, modified algorithms can be applied to terabytes and more of data.

1 Introduction

There is a deluge of electronic data currently available and more coming available. This data comes from diverse sources such as Astronomy where PAN-Starrs is expected to generate 2.5 petabytes of data per year, daily collections of text from newspapers, blogs, etc., medical data that is being collected digitally (including medical images), images of underwater plankton, and more. There will be no class labels for most of this data. For some data sets all of the data will be unlabeled. To explore and make sense of the data, we need approaches such as clustering which groups data into clusters of like data without requiring any class labels [8, 1].

Lawrence O. Hall
Department of Computer Science and Engineering, ENB118,
University of South Florida, Tampa, FL 33620-9951, USA
e-mail: hall@cse.usf.edu

R. Kruse et al. (Eds.): Synergies of Soft Computing and Statistics, AISC 190, pp. 11–15.
springerlink.com

Very large data sets can be practically defined as any which exceed the size of a computers memory. Once data exceeds the size of memory, algorithms to cluster or group the data will run very, very slowly. There will be lots of accesses to secondary storage, typically a disk drive, that will cause the computation time to be bounded by the transfer speed of the disk. Unfortunately, this time is orders of magnitude slower than the speed at which the CPU operates. Hence, we need to have clustering algorithms that are designed for large data and minimize the number of disk accesses. Ideally, all data is loaded into memory only one time. This will mean data is discarded and summaries of it must be retained if all the data is accessed.

In this paper, we discuss variations of fuzzy c-means [7] (FCM) which is an iterative clustering algorithm based on the venerable k-means clustering algorithm [8]. The FCM algorithm is known to converge to a local minima or saddle point [4] as do the variants of it discussed here [3]. In addition to modified clustering algorithms to handle very large data sets, subsampling the data can be explored. The subsample can be random or use intelligent selection. The subsample needs to be smaller than the size of memory to avoid significant slowdowns. Intelligently selecting samples from a very large data set is often infeasible because one likely needs to examine all of the examples, potentially accessing some multiple times.

In this paper, we will discuss the advantages, disadvantages and performance of two large-scale fuzzy clustering algorithms and a subsampling approach that can be used with fuzzy c-means.

2 Single Pass and Online Fuzzy C-Means

One way to process all the data of a very large data set is to apply divide and conquer principles. Load as much data as will fit in memory, cluster it, and keep some type of model of the data for future use. We present two extensions to FCM which work on subsets of the data in different ways. They both rely on weighted examples. We can view each example $\mathbf{x}$ as a feature vector of dimension $\mathbf{s}$. The default weight for an example w_i is 1. FCM variants produce cluster centers or centroids which are representative of the cluster. These centroids can be assigned weights based on the fuzzy memberships of examples in the clusters. So, a cluster would have the weight calculated as shown in Equation 1:

$$w_{c_i} = \sum_{j=1}^{n} u_{ij}, \tag{1}$$

where u_{ij} is the membership of example $\mathbf{x}_j$ in the i^{th} cluster c_i. We can then use the weighted examples representing cluster centers in the clustering

process [2, 10]. The clustering is done with a weighted version of the classic fuzzy c-means algorithm.

The single pass fuzzy c-means clustering (SPFCM) [6] algorithm clusters a chunk of data, creates weighted cluster centers and then processes another chunk of data together with the weighted cluster centers from the previous chunk. The process continues until all the data has been processed. It makes one pass through the data and outputs a set of cluster centers. To determine the cluster each example belongs to a separate pass through the data may be necessary. For newly encountered examples, just a comparison to the final cluster centers is needed.

The online fuzzy c-means (OFCM) clustering algorithm [5] clusters a chunk of data that will fit in memory and obtains weighted cluster centers. It stores the weighted cluster centers. At the end of processing (or intermediate steps) it clusters all of the weighted cluster centers to get a final set of cluster centers. It is designed for streaming (i.e. never ending) data. It can also be applied to existing data sets, where it has some potential advantages.

The SPFCM algorithm will perform poorly if the data is ordered by class, for example. In that case, some chunks will likely have examples from only one, or certainly less than all, class(es) in the data. Since the number of clusters is fixed, this will likely cause a poor set of cluster centers to be created. When there is more than one heavily weighted example from the same cluster it tends to result in multiple final clusters from the same class. Alternatively, the weighted example is assigned to a cluster that represents another class where it will have a strong negative effect on the cluster center location. The problem is exacerbated when clusters are close together.

On the other hand the OFCM clustering algorithm will simply have multiple clusters that can later be combined into one cluster, as long as we choose a fixed number of clusters greater than or equal to the true number. So, in theory it should have less problems when the data for a chunk does not contain a set of examples reflective of the true class distribution.

When applied to existing data, OFCM can be run in parallel on as many processors as necessary (if available). Then the resulting weighted clusters can be clustered. So, it can be completed with just 2 sequential applications of the clustering algorithm, thus allowing for it to be fast in a parallel processing environment.

3 Subsampling

One way of doing subsampling is to select random examples until the subset passes a test [10]. Extensible Fast Fuzzy c-means (eFFCM) randomly samples the dataset (with replacement) in an effort to obtain a statistically significant sample. Statistical significance is tested for with the Chi-square (χ^2) statistic or divergence. If the initial sample fails testing, additional data is added to

the sample until the statistical test is passed [9]. There is time required to do the necessary statistical tests on the random data. However, there is only one data set needed for fuzzy c-means clustering (assuming the size is less than available memory, if not SPFCM or OFCM can be applied).

4 Experiments and Results

One way to evaluate the large data clustering approaches discussed here is to evaluate how close their final cluster centers are to FCM applied to all the data. If they are close or the same, then the speed-ups become important. The algorithms have mostly been evaluated on large volumes of magnetic resonance images of the human brain [6]. The images have 3 features and between 3.8 and 4.3 million examples. The fast algorithm results generally have very good fidelity to FCM. The fastest algorithm is usually SPFCM (with a between 3 and 8 times speed-up depending on chunk size) and eFCM also has a good speed-up and sometimes the best fidelity.

5 Conclusions

Mofications to fuzzy c-means can be used to effectively cluster very large data sets. They have the advantage of convergence and inheriting the well understood properties of FCM. OFCM needs to run in parallel to get the most speed advantages on existing data. SPFCM and eFCM are effective on large data sets with perhaps a little more speed-up using SPFCM. There is a need for scalable clustering algorithms that can find very small clusters and that is a challenge the FCM variants may not be up to.

References

1. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York City (1981)
2. Gu, Y., Hall, L.O., Goldgof, D.B.: Evaluating scalable fuzzy clustering. In: Proc. 2010 IEEE Int. Conf. on Systems Man and Cybernetics (SMC), Istanbul, Turkey, October 10-13, pp. 873–880. IEEE Press (2010)
3. Hall, L., Goldgof, D.: Convergence of the single-pass and online fuzzy c-means algorithms. IEEE Trans. Fuzzy Syst. 19(4), 792–794 (2011)
4. Hathaway, R.J., Bezdek, J.C., Tucker, W.T.: An improved convergence theory for the fuzzy c-means clustering algorithms. In: Bezdek, J.C. (ed.) Analysis of Fuzzy Information: Applications in Engineering and Science, vol. 3, pp. 123–131. CRC Press, Boca Raton (1987)

5. Hore, P., Hall, L., Goldgof, D., Cheng, W.: Online fuzzy c means. In: Ann. Meeting of the North American Fuzzy Information Processing Society (NAFIPS 2008), pp. 1–5 (2008)
6. Hore, P., Hall, L.O., Goldgof, D.B., Gu, Y., Maudsley, A.A., Darkazanli, A.: A scalable framework for segmenting magnetic resonance images. J. Sign. Process. Syst. 54, 183–203 (2009)
7. Hung, M.C., Yang, D.L.: An efficient fuzzy c-means clustering algorithm. In: Proc. 2001 IEEE Int. Conf. on Data Mining (ICDM 2001), pp. 225–232. IEEE Press (2001)
8. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recogn. Lett. 31(8), 651–666 (2010)
9. Pal, N.R., Bezdek, J.C.: Complexity reduction for "large image" processing. IEEE Trans. Syst. Man Cybern. 32(5), 598–611 (2002)
10. Parker, J.K., Hall, L.O., Bezdek, J.C.: Comparison of scalable fuzzy clustering methods. In: Proc. IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE 2012), Brisbane, Australia, June 10-15, pp. 359–367. IEEE Press (2012)

On Creative Uses of Word Associations

Hannu Toivonen, Oskar Gross, Jukka M. Toivanen, and Alessandro Valitutti

Abstract. The ability to associate concepts is an important factor of creativity. We investigate the power of simple word co-occurrence analysis in tasks requiring verbal creativity. We first consider the Remote Associates Test, a psychometric measure of creativity. It turns out to be very easy for computers with access to statistics from a large corpus. Next, we address generation of poetry, an act with much more complex creative aspects. We outline methods that can produce surprisingly good poems based on existing linguistic corpora but otherwise minimal amounts of knowledge about language or poetry. The success of these simple methods suggests that corpus-based approaches can be powerful tools for computational support of creativity.

1 Introduction

The ability to associate concepts, ideas, and problems is an important factor of creativity. Creative people often are able to see or establish connections and analogies where others could not, and this ability may lead to better solutions to problems or new pieces of art.

We are interested in using computers to support or even accomplish tasks involving verbal creativity. In this paper, we will more specifically look at methods that use word associations derived from word co-occurrences in large corpora. For instance, words 'hand' and 'fist' occur relatively often together, indicating that they are semantically related.

More specifically, our goal is to explore the power of word co-occurrences on tasks that require lexical creativity. We keep all other linguistic and world knowledge at a minimum to test how far plain word associations can take

Hannu Toivonen · Oskar Gross · Jukka M. Toivanen · Alessandro Valitutti
Department of Computer Science and HIIT, University of Helsinki, Finland
e-mail: firstname.lastname@cs.helsinki.fi

R. Kruse et al. (Eds.): Synergies of Soft Computing and Statistics, AISC 190, pp. 17–24.
springerlink.com

us. On the other hand, the methods are less dependent on any particular language and resources.

We address two specific tasks. The first one is the taking the Remote Associates Test [10], a psychometric test of creativity. It directly measures the ability to associate words. The second task is generation of poetry, an act with much more complex creative aspects. Also in this case, word associations can be used as a key component of a poetry generation system.

This paper is structured as follows. We first review some background in Section 2. We then address the Remote Associates Test of creativity in Section 3 and generation of poems in Section 4. We conclude in Section 5.

2 Background

We next provide a brief background for word associations: first the RAT creativity test and then word co-occurrence measures.

Remote Associates Test. The Remote Associates Test (RAT) measures the test subject's ability to find associations between words. In the test, three unrelated *cue words* are presented to the subject, e.g., 'thread', 'pine', and 'pain'. The person then tries to identify a fourth word, the *answer word*, which is related to each of the cue words. In this example, the solution is 'needle'.

The Remote Associates Test was developed by Mednick [10] in the 1960s to test creativity defined as "*the forming of associative elements into new combinations, which either meet specified requirements or are in some way useful*". The test is frequently used by psychologists even if some argue that it is not a good measure of creativity.

In practice, RAT measures the ability to discover new associations between concepts that are not typically connected. Performance on RAT also relates to how well one can generate original ideas [5].

Log-likelihood Ratio. We now describe how we use log-likelihood ratio (LLR) to measure how strongly two words are related in a give corpus. We assume a corpus of unstructured documents, and we treat documents as bags of sentences and sentences as bags of words. Instead of sentences, we can consider all n-grams, i.e., sequences of n consecutive words.

The LLR as applied here is based on a multinomial model of co-occurrences of words (see, e.g., Dunning [4]). The multinomial model of any pair $\{x, y\}$ of words has four parameters $p_{11}, p_{12}, p_{21}, p_{22}$, corresponding to the probabilities of events $\{x, y\}$, $\{\neg x, y\}$ $\{x, \neg y\}$ $\{\neg x, \neg y\}$. The ratio of likelihoods of two multinomial models is computed, a null model and an alternative model. The null model assumes independence of words x and y. Their probabilities are estimated as their frequencies in the data, and the probabilities of their different combinations $(p_{11}, \ldots, p_{22})$ are obtained by simple multiplication

(assuming independence). The alternative model, in turn, is the maximum likelihood model which assigns all four parameters from their observed frequencies.

The log-likelihood ratio test is then defined as

$$LLR(x, y) = -2 \sum_{i=1}^{2} \sum_{j=1}^{2} k_{ij} \log(p_{ij}^{null} / p_{ij}), \quad (1)$$

where k_{ij} are the respective counts. LLR measures how much the observed joint distribution of words x and y differs from their distribution under the null hypothesis of independence, i.e., how strong the association between them is in the given corpus.

Related work. Literature on measuring co-occurrences or collocations of words is abundant. Standard techniques include the following.

Log-likelihood ratio is a non-parametric statistical test often used for co-occurrence analysis [4]. Unlike some other measures, log-likelihood ratio does not overestimate the importance of very frequent words.

Latent Semantic Analysis [3] aims to find a set of concepts (instead of terms) in a corpus using singular value decomposition. The semantic similarity (relatedness) of two words can then be estimated by comparing them in the concept space. Latent semantic analysis has then evolved to *Probabilistic Latent Semantic Analysis* [8] and later to *Latent Dirichlet Allocation* [1].

We are also interested in building *networks of word associations*. Concepts maps, mind maps, and mental maps are some well-known examples of specific types of networks designed to help learning and creativity or to model subjective information processing. As an example of work in this area, Tseng et al. [15] proposed a two-phase concept map construction algorithm which uses fuzzy sets and multiple types of rules to generate concept maps.

3 Solving the Remote Associates Test of Creativity

We now illustrate the power of simple word co-occurrence analysis for the RAT test of creativity [7]. This is, admittedly, a narrow and specific context. However, if the human capability to perform well in RAT is related to creativity, then certainly the capability of a computer performing well is an encouraging indication of its ability to potentially perform creative tasks, or at least to help humans in tasks requiring creativity. The more complex task of creating poetry will be addressed in the next section.

Data. We used 212 RAT items of Bowers *et al.* [2] and Mednick & Mednick [11], divided to a training set of 140 items and a test set of 72 items. As a corpus, we use Google 2-grams [12]. We removed stopwords, i.e., common and therefore uninformative English words, using the NLTK stopword list.

3.1 Modeling RAT Items Computationally

Let quadruple $r = (c_1, c_2, c_3, a)$ denote a RAT item, where c_i is the ith cue word and a is the answer word. In a probabilistic formulation, the task is to predict the most likely answer word a given cue words c_1, c_2, c_3. Assuming independence between the cue words, i.e., using the Naïve Bayes model, we obtain

$$P(a|c_1, c_2, c_3) \propto P(a, c_1, c_2, c_3) = P(a) \prod_{i=1}^{3} P(c_i|a). \quad (2)$$

We estimate the (conditional) probabilities from the relative frequencies of the words in the Google 2-grams, and find the word a that maximizes Eq. 2. For more details, see Gross *et al.* [7].

The problem is challenging. There are millions of words to choose from, and even when only considering words that co-occur with each of the cue words, there are thousands of possibilities.

3.2 Experiments

When tested on the RAT items from psychometric literature, the above model provided the correct answer in 66% of cases both in the training and the test sets. Clearly, computers can perform well in such limited tests of creativity by simple co-occurrence analysis even if the search space is very large.

Looking at the 33% of unsuccessful cases, the system often answered with a plural form when the correct answer was singular. Additionally, in some of the test items, a cue word does not occur in the 2-grams at all as an individual word, but only as part of a compound word (with the answer word, for instance). Obviously, one could engineer the method to deal with such issues with plurals and compound words, but the main point is already clear: the performance of the system is better than that of an average person. Item-wise solution rates are typically 30–70%, so the performance of 66% correct solutions can actually be considered very good. This is a clear indication that computers can solve some tasks that are considered to require creativity.

4 Creation of Poetry

We now move on to a much more demanding creative task, writing of poems. We outline a corpus-based approach for this task; more details are given by Toivanen *et al.* [14].

In the literature, several different methods and systems have been proposed for poetry generation (e.g., [9, 6, 16, 13]). They use, among others, statistical

approaches, case-based reasoning, and evolutionary algorithms. Many of the best performing systems are based on explicitly coded knowledge about the world (e.g., using formal logic) as well as rich linguistic knowledge (e.g., a generative grammar or a tagged corpus of poetical text fragments). A different family of approaches is based on Markov chains or n-grams. They learn a model of word sequences from a given corpus and use this model to produce new poetry. The typical shortcoming of such approaches is that longer sequences of text make no sense grammatically or semantically.

Our goal is to minimize all explicit knowledge about the world or the language, and instead rely on given corpora for implicit knowledge about them. Additionally, some off-the-shelf linguistic analysis tools are needed (lemmatizer, part-of-speech tagger, morphological analyzer and synthesizer). We take corpora as input, just like Markov models, but the method is completely different.

We use two corpora. The first one, called *background corpus*, is used to analyze word co-occurrences and to construct a word association network. This network is used to control the topic and semantic coherence of poetry. The second corpus, called *grammar corpus*, is used as a set of grammatical examples or templates in an instance-based manner.

Data. We currently generate poetry in Finnish. The background corpus is Finnish Wikipedia, and the grammar corpus consists of older Finnish poetry.

4.1 Method

The input to the method essentially consists of three items: the background corpus, the grammar corpus, and a topic word.

The contents and coherence of the poem are controlled by using words that are related to the given topic word in the background corpus, as measured by LLR. The grammatical correctness, in turn, is partially guaranteed by taking a random fragment (e.g., a sentence or a poem) from the grammar corpus, and using its grammatical structure in the generated poem.

More specifically, an example fragment of the desired length is chosen from the grammar corpus. It is then analyzed morphologically for the part of speech, case, verb tense, clitics, etc. of each word.

Then, words (especially verbs, nouns, adjectives and adverbs) in the fragment are substituted independently, one by one, by words associated with the given topic. The substitutes are of the same type with the original words and are transformed to similar morphological forms. The original word is left intact, however, if there are no words associated with the topic that can be transformed to the correct morphological form. This can happen, e.g., if the morphological form is rare or complex.

4.2 Results

We next give some example poems generated by the method, translated from Finnish originals. The first poem is about children's play (in the left column). The original text on which it is based (in the right column) is a fragment of a poem by Uuno Kailas.

Computer-generated poem	Text used as a template
How she played then	how she played once
in a daring, daring whispering	in a big green park
under the pale trees.	under the lovely trees.
She had heard for fun	She had watched for fun
how her whispering	how her smile
drifted as jingle to the wind.	fell down as flowers,

The following poem is about hand. The poem fragment used as a template is by Edith Södergran.

Computer-generated poem	Text used as a template
In a pale fist	In a gloomy forest
in a well-balanced fist,	In a dim forest
the buds are so pale	flowers are so pale
in your image lies a dear child god.	In the shadow lies a sick god

The last example is about snow. The text used as a template is by Eino Leino.

Computer-generated poem	Text used as a template
Lives got the frolic ways,	Waves fared the wind's ways,
snow the home of time,	sun the track of time,
softly chimed abandoned homes,	slowly skied for long days,
softly got frolics beloved –	slowly crept for long nights –
ripening crop got the snows' joys.	day wove the deeds of moons

We evaluated the poetry using a panel of twenty random subjects. Each of them evaluated 22 poems, of which 11 were computer-generated and 11 human-written. The poems were presented in a random order and the subjects were not informed that some of the poems are computer-generated. Each poem was evaluated qualitatively along six dimensions: (1) How typical is the text as a poem? (2) How understandable is it? (3) How good is the language? (4) Does the text evoke mental images? (5) Does the text evoke emotions? (6) How much does the subject like the text? These dimensions were evaluated on the scale from one (very poor) to five (very good).

On each of the dimensions, the 67% confidence intervals of the answers for computer-generated vs. human-written poetry overlap a lot (Figure 1), indicating that a large fraction of computer-generated poetry is as good as human-written poetry, even if on average human-written poetry is better.

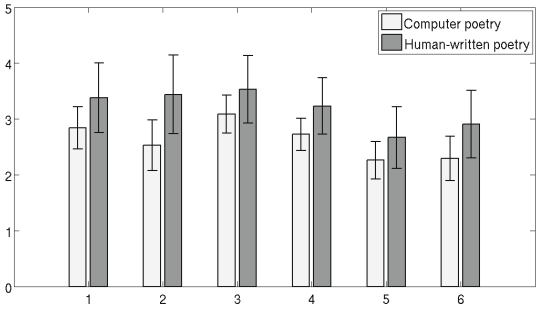

Fig. 1 Subjective evaluation of computer-generated and human-written poetry along six dimensions (see text). Results are averaged over all subjects and poems; whiskers include one standard deviation above and below the mean.

This is a striking result given the simplicity of the methods, and again indicates that simple text analysis methods can be powerful components of verbally creative systems.

5 Conclusions

We have shown how word co-occurrence analysis can be used to perform acts requiring verbal creativity. The Remote Associates Test directly measures the capability to associate words, which is a relatively easy task for a computer when it is given a large corpus. Generation of poetry is a much more complex problem, but word associations together with existing poetry as templates can give surprisingly good results.

The results indicate that word co-occurrence analysis can be a powerful building block of creative systems or systems that support human creativity. While 2-grams were sufficient for achieving a high score on RAT, more relaxed co-occurrences are likely to provide more interesting semantic associations to support or inspire creativity, as suggested by Gross *et al.* [7].

We have used statistical, co-occurrence-based associations of words. The benefit is that their coverage is large, but at the same time they lack explicit semantics. Our results on computational generation of poetry [14] show that this does not prevent them from being used in tasks that demand higher verbal creativity.

In this paper, we have only touched on some specific problems in verbal creativity. We believe that corpus-based approaches can be powerful for many other creative problems, too: they are adaptive and the methods are largely independent of language and resources such as lexicons or knowledge-bases.

Acknowledgements. This work has been supported by the Algorithmic Data Analysis (Algodan) Centre of Excellence of the Academy of Finland.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
2. Bowers, K.S., Regehr, G., Balthazard, C., Parker, K.: Intuition in the context of discovery. Cognitive Psychol. 22(1), 72–110 (1990)
3. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inform. Sci. 41(6), 391–407 (1990)
4. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Comput. Linguist. 19(1), 61–74 (1993)
5. Forbach, G.B., Evans, R.G.: The remote associates test as a predictor of productivity in brainstorming groups. Applied Psychological Measurement 5(3), 333–339 (1981)
6. Gervás, P.: An expert system for the composition of formal spanish poetry. Knowl-Based Syst. 14(3-4), 181–188 (2001)
7. Gross, O., Toivonen, H., Toivanen, J.M., Valitutti, A.: Lexical creativity from word associations. In: Proc. 7th Int. Conf. on Knowledge, Information and Creativity Support Systems (KICSS 2012), Melbourne, Australia, November 8-10. IEEE Press (accepted for publication, 2012)
8. Hofmann, T.: Probabilistic latent semantic indexing. In: ACM SIGIR Conf. on Research and Development in Information Retrieval, SIGIR 1999, pp. 50–57. ACM, New York (1999)
9. Manurung, H., Ritchie, G., Thompson, H.: Towards a computational model of poetry generation. In: Proc. AISB 2000 Symposium on Creative and Cultural Aspects and Applications of AI, pp. 79–86 (2000)
10. Mednick, S.: The associative basis of the creative process. Psychol. Rev. 69(3), 220–232 (1962)
11. Mednick, S., Mednick, M.: Remote associates test, examiner's manual: college and adult forms 1 and 2, Boston, MA, USA (1967)
12. Michel, J., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., Aiden, E.L.: Quantitative analysis of culture using millions of digitized books. Science 331(6014), 176–182 (2011)
13. Netzer, Y., Gabay, D., Goldberg, Y., Elhadad, M.: Gaiku: generating haiku with word associations norms. In: Proc. Workshop on Computational Approaches to Linguistic Creativity (CALC 2009), pp. 32–39. Association for Computational Linguistics, Stroudsburg (2009)
14. Toivanen, J.M., Toivonen, H., Valitutti, A., Gross, O.: Corpus-based generation of content and form in poetry. In: Proc. Int. Conf. on Computational Creativity (ICCC 2012), Dublin, Ireland, May 30-June 1, pp. 175–179 (2012)
15. Tseng, S., Sue, P., Su, J., Weng, J., Tsai, W.: A new approach for constructing the concept map. Comput. Educ. 49(3), 691–707 (2007)
16. Wong, W.T., Chun, A.H.W.: Automatic haiku generation using VSM. In: Proc. Int. Conf. on Advances on Applied Computer and Applied Computational Science (ACACOS 2008), pp. 318–323 (2008)

Part II
Foundations

Combining Imprecise Probability Masses with Maximal Coherent Subsets: Application to Ensemble Classification

Sébastien Destercke and Violaine Antoine

Abstract. When working with sets of probabilities, basic information fusion operators quickly reach their limits: intersection becomes empty, while union results in a poorly informative model. An attractive means to overcome these limitations is to use maximal coherent subsets (MCS). However, identifying the maximal coherent subsets is generally NP-hard. Previous proposals advocating the use of MCS to merge probability sets have not provided efficient ways to perform this task. In this paper, we propose an efficient approach to do such a merging between imprecise probability masses, a popular model of probability sets, and test it on an ensemble classification problem.

Keywords: Ensemble, inconsistency, information fusion, maximal coherent subsets.

1 Introduction

When multiple sources provide information about the ill-known value of some variable X it is necessary to aggregate these pieces of information into a single model. In the case where the initial uncertainty models are precise probabilities and where the aggregated model is constrained to be precise as well, there are only a few options to combine the information (see [3] for a complete review).

The situation changes when one considers imprecision-tolerant uncertainty theories, such as possibility theory, evidence theory or imprecise

Sébastien Destercke
UMR Heudiasyc, Université de Technologie de Compiègne,
60200 Compiègne, France
e-mail: sebastien.destercke@hds.utc.fr

Violaine Antoine
LISTIC, Polytech Annecy-Chambéry, 74944 Annecy le Vieux, France
e-mail: violaine.antoine@univ-savoie.fr

R. Kruse et al. (Eds.): Synergies of Soft Computing and Statistics, AISC 190, pp. 27–35.
springerlink.com

probability theory (see [7]). As they extend both set-theoretic and probabilistic approaches[1], these theories can use aggregation operators coming from both frameworks, i.e., they can generalise intersections and unions of sets as well as averaging methods.

When there is (strong) conflict between information pieces, both conjunctive (intersection) and disjunctive (union) aggregation face some problems: conjunction results are often empty and disjunction results are often too imprecise to be really useful. A theoretically attractive solution to these problems is to use maximal coherent subsets (MCS) [11], that is to consider subsets of sources who are consistent and that are maximal with this property. Aggregation can then be done by combining conjunction within maximal coherent subsets with other aggregation operators, e.g., disjunction. Practically, the main difficulty that faces this approach is to identify MCS, a NP-hard problem in the general case.

Different solutions have been proposed to combine inconsistent pieces of information within the framework of imprecise probability theory. In [10] and [12], hierarchical models are considered. In [1] and [8], Bayesian-like methods (i.e., using conditional probabilities) of aggregation are proposed. In [9] and [14], non-Bayesian methods are studied (although [14] considers that combination methods should commute with Bayesian updating). In the two latter references, MCS are proposed as a solution to combine information pieces that are partially inconsistent, but no practical methods are given to identify MCS.

In this paper, we concentrate on imprecise probability masses and propose a practical approach to apply MCS inspired combination methods to such models. We work in a non-Bayesian framework. Section 2 recalls the necessary background on imprecise probabilities and information fusion. Section 3 describes our approach, of which the most important part is the algorithm to identify MCS. Finally, Section 4 presents an application to ensemble classification, in which resulting classification models are combined using MCS.

2 Preliminaries

The theory of imprecise probabilities [15] is a highly expressive framework to represent uncertainty. This section presents the basics of imprecise probabilities.

2.1 Imprecise Probabilities

Consider a variable X taking values in a finite domaine D_x of n elements $\{x_1, x_2, \ldots, x_n\}$. Basically, imprecise probabilities characterize uncertainty about X by a closed convex set $\mathcal{P}$ of probabilities defined on D_x. To this

[1] Except possibility theory, that does not encompass probabilities as special cases.

set $\mathcal{P}$ can be associated Lower and upper probabilities that are mappings from the power set 2^{D_x} to $[0,1]$. They are respectively denoted $\underline{P}$ and $\overline{P}$ and are defined, for an event $A \subseteq D_x$, as $\underline{P}(A) = \inf_{p\in\mathcal{P}} P(A)$ and $\overline{P}(A) = \sup_{p\in\mathcal{P}} P(A)$. These two measures are dual, in the sense that $\underline{P}(A) = 1 - \overline{P}(A^c)$, with A^c the complement of A. Hence, all the information is contained in only one of them.

Alternatively, one can start from a lower measure $\underline{P}$ and compute the convex set $\mathcal{P}_{\underline{P}} = \{P \in \mathbb{P}(D_x) | P(A) \geq \underline{P}(A),\ \forall A \subseteq D_x\}$ of dominating probability measures ($\mathbb{P}(D_x)$ is the set of all probabilities on D_x). Note that the lower value $P_*(A) = \inf_{P\in\mathcal{P}_{\underline{P}}} P(A)$ need not coincide with $\underline{P}(A)$ in general. If the equality $P_* = \underline{P}$ holds, then $\underline{P}$ is said to be *coherent*. In this paper, we will deal exclusively with coherent lower probabilities. Note that lower probabilities are not sufficient to represent every possible convex sets of probabilities.To represent any convex set $\mathcal{P}$, one actually needs to consider bounds on expectations (see [15]).

2.2 Imprecise Probability Masses (IPM)

Usually, the handling of generic sets $\mathcal{P}$ (and sets represented by lower probabilities) represent a heavy computational burden. In practice using simpler models alleviate this computational burden to the cost of a lower expressivity. Imprecise probability masses [4] (IPM) are such simpler models.

IPM can be represented as a family of intervals $L = \{[l_i, u_i], i = 1, \ldots, n\}$ verifying $0 \leq l_i \leq u_i \leq 1 \forall i$. The interval bounds are interpreted as probability bounds over singletons. They induce a set $\mathcal{P}_L = \{p \in \mathbb{P}(D_x) | l_i \geq p(x_i) \geq u_i,\ \forall x_i \in D_x\}$. An extensive study of IPM and their properties can be found in [4].

A set L of IPM is said to be proper if the condition $\sum_{i=1}^n l_i \geq 1 \geq \sum_{i=1}^n u_i$ holds, and $\mathcal{P}_L \neq \emptyset$ if and only if L is proper. Considered sets are always proper, other types having no interest. To guarantee that lower and upper bounds are reachable for each singleton x_i by at least one probability in $\mathcal{P}_L$, the intervals must verify:

$$\sum_{i\neq j} l_j + u_i \leq 1 \text{ and } \sum_{i\neq j} u_j + l_i \geq 1 \quad \forall i. \tag{1}$$

If L is reachable, lower and upper probabilities of $\mathcal{P}_L$ can be computed as follows:

$$\underline{P}(A) = \max(\sum_{x_i\in A} l_i, 1 - \sum_{x_i\notin A} u_i), \quad \overline{P}(A) = \min(\sum_{x_i\in A} u_i, 1 - \sum_{x_i\notin A} l_i). \tag{2}$$

If L is not reachable, a reachable set L' is obtained by applying Eq. (2) to singletons.

2.3 Basic Combinations of Imprecise Probabilities

When M sources provide information, there are three basic ways to combine this information: through a conjunction, a disjunction or a weighted mean. When information is given by credal sets $\mathcal{P}_i$, $i = 1, \ldots, M$, computing these basic combination results present some computational difficulties [9]. Computations become much easier if we consider a set $L_1, \ldots, L_M$ of IPM. In this case, if $l_{i,j}, u_{i,j}$ denote respectively the lower and upper probability bounds on element x_i given by source j, (approximated) combinations are as follows:

- Weighted mean ($L_\sum$): $l_{i,\sum} = \sum_{j=1,M} w_j l_{i,j}, \quad u_{i,\sum} = \sum_{j=1,M} w_j u_{i,j}$
- Disjunction ($L_\cup$): $l_{i,\cup} = \min_{j=1,M} l_{i,j}, \quad u_{i,\cup} = \max_{j=1,M} u_{i,j}$
- Conjunction ($L_\cap$):

$$l_{i,\cap} = \max(\max_{j=1,M} l_{i,j}, 1 - \sum_{k \neq i} \min_{j=1,M} u_{i,j}), \tag{3}$$

$$u_{i,\cap} = \min(\min_{j=1,M} u_{i,j}, 1 - \sum_{k \neq i} \max_{i=1,M} l_{i,j})$$

In general, the bounds obtained by conjunction (3) may be non-proper, i.e. may result in an empty $\mathcal{P}_{L_\cap}$. $L_1, \ldots, L_M$ have a non-empty intersection iff the following conditions [4] hold:

$$\max_{j=1,M} l_{i,j} \leq \min_{j=1,M} u_{i,j} \text{ for every } i \in [1, n] \tag{4}$$

$$\sum_{i=1,n} \max_{j=1,M} l_{i,j} \leq 1 \leq \sum_{i=1,n} \min_{j=1,M} u_{i,j} \tag{5}$$

The first condition ensures that intervals have a non-empty intersection for every singleton, while the second makes sure that the result is a proper probability interval.

3 Maximal Coherent Subsets (MCS) and IPM

This section describes the methods to identify and combine MCS.

3.1 Identifying MCS

When sources provide sets $\mathcal{P}_1, \ldots, \mathcal{P}_M$, finding MCS comes down to find every subset $K \subseteq [1, M]$ such that $\bigcap_{i \in K} \mathcal{P}_i \neq \emptyset$ and such that K is maximum with this property (i.e., adding a new set would make the intersection empty).

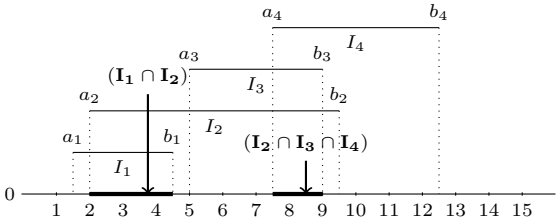

Fig. 1 Maximal coherent subsets on Intervals

Usually, identifying every possible coherent subset among $\mathcal{P}_1, \ldots, \mathcal{P}_M$ is NP-hard, making it a difficult problem to solve in practice.

A particularly interesting case where MCS can be found easily is when each sources provide intervals $[a_i, b_i]$, $i = 1, \ldots, M$. In this case, Algorithm 1 given in [6] finds MCS. It requires to sort values $\{a_i, b_i | i = 1, \ldots, M\}$ (complexity in $\mathcal{O}(M \log M)$), and is then linear in the number of sources.

Algorithm 1. Maximal coherent subsets of intervals

Input: M intervals
Output: List of S maximal coherent subsets K_j

```
1  List = ∅, j=1, K = ∅ ;
2  Order in an increasing order {a_i|i = 1,...,M} ∪ {b_i|i = 1,...,M} ;
3  Rename them {c_i|i = 1,...,2M} with type(i) = a if c_i = a_k and type(i) = b if
   c_i = b_k ;
4  for i = 1,...,2M − 1 do
5      if type(i) = a then
6          Add Source k to K s.t. c_i = a_k ;
7          if type(i + 1) = b then
8              Add K to List (K_j = K) ;
9              j = j + 1 ;
10     else
11         Remove Source k from K s.t. c_i = b_k ;
```

This algorithm can be applied directly to IPM intervals to check MCS satisfying Condition (4) (which is necessary for a subset of IPM to have a non-empty intersection). Indeed, consider a singleton x_i and the set of intervals $L^i = [l_{i,j}, u_{i,j}]$, $j = 1, M$: if $K \subseteq [1, M]$ is not a MCS of L^i, then the credal sets $\{\mathcal{P}_j | j \in K\}$ do not form a MCS. Hence, iteratively applying Algorithm 1 as exposed in Algorithm 2 allows to easily identify possible MCS among sets $\mathcal{P}_{L_1}, \ldots, \mathcal{P}_{L_M}$. In each iteration (Line 2), Algorithm 2 refines the MCS found in the previous one (stored in *List*) by finding MCS for probaiblity intervals of singleton x_i (Line 5).

Algorithm 2. MCS identification for IPM

Input: M IPM
Output: List of S possible maximal coherent subsets K_j

```
1 List = {{1, M}} ;
2 for i = 1, ..., n do
3     K=∅;
4     foreach subset E in List do
5         Run Algorithm 1 on [l_{i,j}, u_{i,j}], j ∈ E ;
6         Add resulting list of MCS to K;
7     List= K;
```

Example 1. Consider the IPM defined on $D_x = \{x_1, x_2, x_3\}$ and summarised in Table 1. Running Algorithm 2 then provides successively the following MCS: $K = \{\{1,2,3\},\{2,3,4\}\}$ after the first iteration ($i = 1$ in Line 2 of Algorihm 2); $K = \{\{1,2\},\{3\},\{2,4\}\}$ after the second iteration, and K is not changed during the third iteration.

Table 1 Examples of IPM

*Source*1				*Source*2				*Source*3				*Source*4			
	x_1	x_2	x_3		x_1	x_2	x_3		x_1	x_2	x_3		x_1	x_2	x_3
$u_{i,1}$	0.6	0.5	0.2	$u_{i,2}$	0.55	0.55	0.2	$u_{i,3}$	0.5	0.2	0.6	$u_{i,4}$	0.35	0.6	0.35
$l_{i,1}$	0.4	0.3	0.	$l_{i,2}$	0.35	0.35	0.	$l_{i,3}$	0.3	0.	0.4	$l_{i,4}$	0.15	0.4	0.15

Some subsets $K_1, \ldots, K_S$ of sources resulting from Algorithm 2 do not satisfy Condition (5). If K_ℓ is such a set, then one can either make an (exponential) exhaustive search of MCS within K_ℓ, or correct IPM in K_ℓ in a minimal way, so that they satisfy (5). In this last case we can transform [9] bounds $l_{i,j}$ and $u_{i,j}$, $j \in K_\ell, i \in [1,n]$ into $l'_{i,j} = \epsilon l_{i,j}$ and $u'_{i,j} = \epsilon u_{i,j} + (1-\epsilon)$ with ϵ the minimal value such that

$$\sum_{i=1,n} \max_{j \in K_\ell} l'_{i,j} \leq 1 \leq \sum_{i=1,n} \min_{j \in K_\ell} u'_{i,j}. \tag{6}$$

This strategy makes the identification of MCS easy. Roughly speaking, it applies Algorithm 1 to probabilistic (expectation) bounds coming from different sources but bearing on common events (functions). Note that the same strategy can be applied to models based on peculiar families of events (functions), such as p-boxes [5].

3.2 Combination with MCS

Once MCS $K_1, \ldots, K_S$ of sources have been identified, they can be used to combine inconsistent information. Without loss of generality, consider the indexing such that $|K_1| \geq \ldots \geq |K_S|$ where $|K_i|$ is the cardinality of K_i (i.e., the number of sources within it).

We then propose two ways of combining the probability sets $\mathcal{P}_{L_1}, \ldots, \mathcal{P}_{L_M}$. In both of them, we consider the IPM L_{K_ℓ}, $\ell = 1, \ldots, S$ obtained by combining IPM in K_ℓ according to the conjunctive rule (3).

The first rule combines disjunctively the first n IPM L_{K_i}, that is

$$l_{i,\cup\cap_n} = \min_{\ell=1,n} l_{i,K_\ell}, \quad u_{i,\cup\cap_n} = \max_{\ell=1,n} u_{i,K_\ell} \tag{7}$$

where $l_{i,K_\ell}, u_{i,K_\ell}$ are the probability bounds given by L_{K_ℓ} on x_i.

The second rule combines by a weighted mean the first n, L_{K_i}, that is

$$l_{i,\cup\cap_n} = \sum_{\ell=1,n} w_{\ell,n} l_{i,K_\ell}, \quad u_{i,\cup\cap_n} = \sum_{\ell=1,n} w_{\ell,n} u_{i,K_\ell} \tag{8}$$

where $w_{\ell,n} = {}^{|K_\ell|}/_{\sum_{i=1,n}|K_i|}$ is the importance of K_ℓ in number of sources (a similar strategy is used in [9]). If $n = S$ the rules simply combine every MCS.

4 Application to Ensemble Classification

Combination is an essential feature of ensemble classification. As classifiers often disagree together, using a MCS based approach to combine the different sources appears sensible. We have therefore tested our approach in the following way: we have trained forest of decision trees; for a given instance and for each decision trees, we have built an IPM model using the Imprecise Dirichlet model (IDM) with an hyperparameter $s = 4$ (see [2] for details) and taking the samples in the tree leaves as observations. We then combined the different IPM with the two rules (7) and (8) (n=5) and selected the final class according to the maximin and maximality criterion (see [13] for details). The former results in a unique decision while the latter results in a set of possible optimal decisions.

Classifier performances are estimated using discounted accuracy: assume we have T observations whose classes $x^i, i = 1, \ldots, T$ are known and for which T (possibly imprecise) predictions $\widehat{X}^1, \ldots, \widehat{X}^T$ have been made. The discounted accuracy $d - acc$ of the classifier is then

$$d - acc = \frac{1}{T} \sum_{i=1}^{T} \frac{\Delta_i}{f(|\widehat{X}_i|)}, \tag{9}$$

with $\Delta_i = 1$ if $x^i \in \widehat{X}^i$, zero otherwise and f an increasing function such that $f(1) = 1$. Set accuracy $(s - acc)$ is obtained with $f(|\widehat{X}_i|) = 1$.

Results are summarized in Table 2. Numbers of trees in the forest are $\{10, 20, 50\}$ and the data sets are Zoo, Segment and Satimage (taken from UCI), all of them with 7 classes. Results were compared to a classical voting strategy. We have also indicated the average CPU time needed to apply the different combination rules. From the results, it appears that using a conjunctive rule between provided imprecise probabilistic models does not improve much the results of classical voting. This is not surprising as we use precise decision trees and IDM to build our models, and it would be worthwhile to check whether these conclusions still hold when using credal classifiers. The interest of using imprecise probabilistic models appears when we allow for some imprecision, that is when we adopt a partially disjunctive rule (Rule (7) with n=5). In this latter case, allowing for imprecise classification increases the percentage of well-recognized instances while not decreasing too much the precision. Finally, we can notice that the average computational time does not increase much when the number of sources increases.

Table 2 Results summary. d-acc: discounted accuracy, s-acc: set accuracy, acc: standard accuracy (with maximin)

data	Tree	Single	Votes	Rule (7) (n=1)			Rule (7) (n=5)			Rule (8) (n=5)			avg CPU
set	nb	tree		d-acc	s-acc	acc	d-acc	s-acc	acc	d-acc	s-acc	acc	time
Sat	10	0.81	0.88	0.87	0.88	0.87	0.64	0.98	0.81	0.87	0.89	0.87	15.16
	20	0.81	0.89	0.88	0.89	0.88	0.61	0.98	0.82	0.88	0.90	0.88	23.80
	50	0.81	0.89	0.89	0.90	0.89	0.63	0.97	0.85	0.89	0.90	0.89	61.28
Zoo	10	0.91	0.92	0.86	0.86	0.86	0.69	0.96	0.80	0.88	0.92	0.92	0.28
	20	0.91	0.93	0.79	0.79	0.79	0.60	0.99	0.68	0.85	0.91	0.88	0.50
	50	0.91	0.93	0.90	0.92	0.91	0.69	0.96	0.87	0.86	0.88	0.88	2.34
Seg	10	0.93	0.96	0.96	0.96	0.96	0.72	1.00	0.84	0.95	0.97	0.96	5.59
	20	0.93	0.95	0.95	0.95	0.95	0.66	0.99	0.81	0.95	0.96	0.95	8.03
	50	0.93	0.96	0.96	0.96	0.96	0.64	0.98	0.84	0.96	0.96	0.96	17.64

5 Conclusion

In this paper, we have proposed an efficient way to find MCS with imprecise probability masses, and have applied it to the combination of multiple classifiers. First results indicate that using a disjunctive approach to combine conjunctively merged MCS may quickly result in poorly informative models, hence it may be safer in general to adopt other strategies (e.g., combining only a limited number of MCS or using a weighted mean).

Note that the algorithms presented here can be applied to other imprecise probabilistic models as well, as long as they are defined by probability bounds bearing on the same events (or by expectation bounds bearing on the same function).

References

1. Benavoli, A., Antonucci, A.: An aggregation framework based on coherent lower previsions: Application to zadeh's paradox and sensor networks. Int. J. Approx. Reason. 51(9), 1014–1028 (2010)
2. Bernard, J.: An introduction to the imprecise Dirichlet model for multinomial data. Int. J. Approx. Reason. 39(2-3), 123–150 (2005)
3. Cooke, R.: Experts in Uncertainty. Oxford University Press, Oxford (1991)
4. de Campos, L., Huete, J., Moral, S.: Probability intervals: a tool for uncertain reasoning. Int. J. Uncertain. Fuzz. 2, 167–196 (1994)
5. Destercke, S., Dubois, D.: The role of generalised p-boxes in imprecise probability models. In: Proc. of the 6th Int. Symp. on Imprecise Probability: Theories and Applications, pp. 179–188 (2009)
6. Dubois, D., Fargier, H., Prade, H.: Multi-source information fusion: a way to cope with incoherences. In: Proc. of French Days on Fuzzy Logic and Applications (LFA), La Rochelle, France, pp. 123–130 (2000)
7. Dubois, D., Prade, H.: Formal representations of uncertainty. In: Decision-Making Process, ch. 3, pp. 85–156. ISTE, London (2010)
8. Karlsson, A., Johansson, R., Andler, S.F.: On the behavior of the robust Bayesian combination operator and the significance of discounting. In: Proc. of the 6th Int. Symp. on Imprecise Probability: Theories and Applications, pp. 259–268 (2009)
9. Moral, S., Sagrado, J.: Aggregation of imprecise probabilities. In: Aggregation and Fusion of Imperfect Information, pp. 162–188. Physica, Heidelberg (1997)
10. Nau, R.: The aggregation of imprecise probabilities. J. Stat. Plan. Infer. 105, 265–282 (2002)
11. Rescher, N., Manor, R.: On inference from inconsistent premises. Theor. Decis. 1, 179–219 (1970)
12. Troffaes, M.: Generalising the conjunction rule for aggregating conflicting expert opinions. Int. J. Intell. Syst. 21(3), 361–380 (2006)
13. Troffaes, M.: Decision making under uncertainty using imprecise probabilities. Int. J. Approx. Reas. 45, 17–29 (2007)
14. Walley, P.: The elicitation and aggregation of beliefs. Technical report, University of Warwick (1982)
15. Walley, P.: Statistical reasoning with imprecise Probabilities. Chapman and Hall, New York City (1991)

The Goodman-Nguyen Relation in Uncertainty Measurement

Renato Pelessoni and Paolo Vicig

Abstract. The Goodman-Nguyen relation generalises the implication (inclusion) relation to conditional events. As such, it induces inequality constraints relevant in extension problems with precise probabilities. We extend this framework to imprecise probability judgements, highlighting the role of this relation in determining the natural extension of lower/upper probabilities defined on certain sets of conditional events. Further, a generalisation of the Goodman–Nguyen relation to conditional random numbers is proposed.

Keywords: Goodman-Nguyen relation, imprecise probabilities.

1 Introduction

It is well known that probability constraints depend essentially on relations among events. In particular, the *monotonicity* requirement that

$$(E \Rightarrow F) \rightarrow \mu(E) \leq \mu(F), \tag{1}$$

μ being a probability or a more general uncertainty measure, is a very minimal one. In fact, $E \Rightarrow F$ means that F is certainly true whenever E is true, but might possibly be true even in cases when E is false: then obviously F must be at least as likely as E. In fact, (1) holds also when μ is a coherent lower/upper probability, or a capacity. In the latter case, it is generally taken as one of the defining properties of capacities.

The implication relation '$\Rightarrow$' also plays a role in the following problem, a special case of de Finetti's Fundamental Theorem [3]: given a coherent

Renato Pelessoni · Paolo Vicig
Università di Trieste, 34127, Trieste, Italy
e-mail: {renato.pelessoni,paolo.vicig}@econ.units.it

R. Kruse et al. (Eds.): Synergies of Soft Computing and Statistics, AISC 190, pp. 37–44.
springerlink.com

probability P on the set $\mathcal{A}(\mathbb{P})$ of all events (logically) dependent on a given partition $\mathbb{P}$, which are its coherent extensions to an additional event $E \notin \mathcal{A}(\mathbb{P})$? The well known answer is that $P(E)$ must be chosen in a closed interval, $P(E_*) \leq P(E) \leq P(E^*)$, with $E_* = \vee\{e \in \mathbb{P} : e \Rightarrow E\}$, $E^* = \vee\{e \in \mathbb{P} : e \wedge E \neq \varnothing\}$, $E_*, E^* \in \mathcal{A}(\mathbb{P})$.

A generalisation of the implication relation to *conditional* events, the Goodman-Nguyen (in short: GN) relation $\leq_{GN}$ was apparently first introduced in [4], and some of its implications for precise conditional probabilities were studied in [1, 2, 5].

The main purpose of this paper is to further explore the relevance of the GN relation in more general cases. Precisely, we consider imprecise conditional probabilities which are either coherent or C-convex. Some preliminary material is recalled in Section 2, including a survey of known facts about the GN relation. We then discuss the generalisation of the basic result (1) to

$$A|B \leq_{GN} C|D \to \mu(A|B) \leq \mu(C|D) \tag{2}$$

in Section 3, and the role of the GN relation in extension problems in Section 3.1. The main results here are Propositions 3 and 4; in particular, Proposition 4 determines the natural extension of a coherent lower probability (alternatively the convex natural extension of a C-convex lower probability) assessed on a structured set $\mathcal{A}_C$ of conditional events. In Section 4 we explore the possibility of further extending the GN relation to conditional random numbers, and hence to employ it with coherent lower previsions. To the best of our knowledge, these questions have not been tackled yet in the relevant literature. By Proposition 5, the generalisation we propose ensures an analogue of eq. (2), whilst it is less clear if and how it may be employed in extension problems. Some final considerations are included in Section 5. Due to space limitations, most proofs have been omitted.

2 Preliminaries

In the sequel, following [3, 4] and others, we employ the logical rather than the set theoretical notation for operations with events. In terms of a truth table, a conditional event $A|B$ can be thought of as true, when A and B are true, false when A is false and B true, undefined when B is false. It ensues that $A|B$ and $A \wedge B|B$ have the same logical values, i.e. $A|B = A \wedge B|B$.

Given a *partition* $\mathbb{P}$, i.e. a set of pairwise disjoint events whose logical sum (union) is the sure event Ω, an event E is *logically dependent* on $\mathbb{P}$ iff E is a logical sum of events of $\mathbb{P}$, $E = \vee\{\omega \in \mathbb{P} : \omega \Rightarrow E\}$. The set $\mathcal{A}(\mathbb{P})$ of all events logically dependent on $\mathbb{P}$ is a field (also called the powerset of $\mathbb{P}$).

The precise or imprecise *conditional* probabilities considered in the sequel will often be defined on $\mathcal{A}_C = \mathcal{A}_C(\mathbb{P}) = \{A|B : A, B \in \mathcal{A}(\mathbb{P}), B \neq \varnothing\}$.

Given $\mathbb{P}$, a conditional random number $X|B$, $B \in \mathcal{A}(\mathbb{P}) - \{\varnothing\}$, takes up the values $X(\omega)$, for $\omega \in \mathbb{P}$, $\omega \Rightarrow B$, is undefined for $\omega \Rightarrow B^c$. When $B = \Omega$, $X|\Omega = X$ is a(n unconditional) random number. The *indicator* I_A of an event A is the simplest non-trivial random number; we shall often denote A and its indicator I_A with the same letter A. Note that $A \Rightarrow B$ is equivalent to $I_A \leq I_B$. An arbitrary set of events $S = \{E_i : i \in I\}$ does generally not constitute a partition, but originates the *partition* $\mathbb{P}_g$ *generated* by S, whose elements are the logical products $\wedge_{i \in I} E_i'$, where for each $i \in I$ the symbol E_i' can be replaced by either event E_i or its negation E_i^c. By specifying E_i' for each i in all possible ways, we get the elements (some of them, in general, may be impossible) of $\mathbb{P}_g$. The events in S belong to $\mathcal{A}(\mathbb{P}_g)$.

A *lower prevision* $\underline{P}$ on a set S of (*bounded*, in what follows) conditional random numbers is a map $\underline{P} : S \longmapsto \mathbb{R}$. If S has the property $X|B \in S \rightarrow -X|B \in S$, its conjugate *upper prevision* is defined as $\overline{P}(X|B) = -\underline{P}(-X|B)$. Because of conjugacy, we may employ lower or alternatively upper previsions only.

Definition 1. A lower prevision $\underline{P} : S \longmapsto \mathbb{R}$ is *W-coherent* iff, for all $n \in \mathbb{N}$, $\forall X_0|B_0, \ldots, X_n|B_n \in S$, $\forall\ s_0, s_1, \ldots, s_n$ real and *non-negative*, defining $B^* = \bigvee_{i=0}^n B_i$ and $\underline{G} = \sum_{i=1}^n s_i B_i (X_i - \underline{P}(X_i|B_i)) - s_0 B_0 (X_0 - \underline{P}(X_0|B_0))$, the following condition holds: $\sup(\underline{G}|B^*) \geq 0$.

This is essentially Williams' definition of coherence [12], as restated in [8], and is equivalent to Walley's definition 7.1.4(b) in [11], if S is made up of a finite number of conditional random numbers, each with finitely many values. If $X|B = X|\Omega = X$, $\forall X|B \in S$, it reduces to Walley's (unconditional) coherence ([11], Sec. 2.5.4 (a)).

A weaker concept than W-coherence is that of *C–convex* conditional lower prevision, obtained from Definition 1 by introducing just the extra *convexity constraint* $\sum_{i=1}^n s_i = s_0\ (> 0)$ and requiring that $\underline{P}(0) = 0$. These previsions were studied in [7] and correspond to certain kinds of risk measures.

Coherent conditional previsions may be defined similarly:

Definition 2. $P : S \longmapsto \mathbb{R}$ is a coherent conditional prevision iff, for all $n \in \mathbb{N}$, $\forall\ X_1|B_1, \ldots, X_n|B_n \in S$, $\forall\ s_i \in \mathbb{R}$ $(i = 1, \ldots, n)$, defining $G = \sum_{i=1}^n s_i B_i (X_i - P(X_i|B_i))$, $B^* = \bigvee_{i=1}^n B_i$, it holds that $\sup(G|B^*) \geq 0$.

In the consistency concepts above, we may speak of (lower, upper or precise) *probability* μ instead of prevision if, for any $X|B \in S$, X is (the indicator of) an event. In all such cases, the following are necessary consistency conditions:

$$\mu(A|B) \in [0; 1], \mu(\varnothing|B) = 0, \mu(B|B) = 1. \tag{3}$$

In general, results for upper probabilities follow from those for lower probabilities by the conjugacy equality $\overline{P}(A) = 1 - \underline{P}(A^c)$.

Definition 3. *(Goodman–Nguyen relation.)* We say that $A|B \leq_{GN} C|D$ iff

$$A \wedge B \Rightarrow C \wedge D \text{ and } C^c \wedge D \Rightarrow A^c \wedge B. \tag{4}$$

The GN relation was introduced in an equivalent form in [4], while investigating *conditional event algebras*. As observed in [1, 5, 6], the intuition behind Definition 3 can be explained easily resorting to betting arguments, much in the style of de Finetti [3]. In fact, (4) states that whenever we bet both on $A|B$ and on $C|D$ (iff $B \wedge D$ is true), if we win the bet on $A|B$, we also win the bet on $C|D$ (because $A \wedge B \Rightarrow C \wedge D$), and losing the bet on $C|D$ implies also our losing the bet on $A|B$ (because of $C^c \wedge D \Rightarrow A^c \wedge B$). When $B = D = \Omega$, just one of the implications in (4) is needed, because of the tautology $A \Rightarrow C \leftrightarrow C^c \Rightarrow A^c$.

Not surprisingly then, (2) should hold. In the case that μ is a conditional probability P, (2) was stated without proof in [4] (assuming P defined on $\mathcal{A}_C$), and proved in [1] (under general assumptions) and independently (in a less general case) in [5]. The result ensues also from Proposition 1 in Section 3, supplying a unique proof for either precise or imprecise probabilities. The GN relation in extension problems is explored in [1], showing that, given a coherent probability P on a *finite* set of conditional events, the bounds on its coherent extensions on one additional event $C|D$ depend on the values of P on two events, $(C|D)_*$ and $(C|D)^*$, determined by the GN relation. We consider extension problems for *arbitrary* sets of events and precise or imprecise probability assessments in Section 3.

3 The GN Relation with Imprecise Probabilities

Remark 1. If $A|B \leq_{GN} C|D$, then necessarily the partition $\mathbb{P}_g$ generated by A, B, C, D allows for at most 7 non-impossible events that imply $B \vee D$: $\omega_1 = ABCD$, $\omega_2 = A^cBCD$, $\omega_3 = AB^cCD$, $\omega_4 = A^cB^cCD$, $\omega_5 = A^cBC^cD$, $\omega_6 = A^cBCD^c$, $\omega_7 = A^cBC^cD^c$. This is easily seen from (4), using $A \Rightarrow B \leftrightarrow A \wedge B^c = \varnothing$.

Example 1. If $A \Rightarrow C \Rightarrow D \Rightarrow B$, then $A|B \leq_{GN} C|D$. Of the 7 events in Remark 1, only 4 (at most, iff $A \neq \varnothing$) are non-impossible: ω_1, ω_2, ω_5, ω_7.

We prove now equation (2) for C-convex probabilities.[1]

Proposition 1. *Let μ be a C-convex lower (or upper) probability defined on $S = \{A|B, C|D\}$. Then, $A|B \leq_{GN} C|D$ implies $\mu(A|B) \leq \mu(C|D)$.*

Proof. Consider, in the lower probability case,

$$\underline{G}|B^* = \underline{G}|B \vee D = B(A - \mu(A|B)) - D(C - \mu(C|D))|B \vee D. \qquad (5)$$

Assuming $A|B \leq_{GN} C|D$, by Remark 1 $\underline{G}|B \vee D$ can take up at most 7 values, actually fewer distinct ones: $\underline{G}(\omega_1) = \underline{G}(\omega_5) = \mu(C|D) - \mu(A|B)$,

[1] Since any W-coherent or coherent probability is C-convex, Proposition 1 applies to these probabilities too.

$\underline{G}(\omega_2) = \mu(C|D) - \mu(A|B) - 1 \leq 0$, $\underline{G}(\omega_3) = \underline{G}(\omega_4) = \mu(C|D) - 1 \leq 0$, $\underline{G}(\omega_6) = \underline{G}(\omega_7) = -\mu(A|B) \leq 0$. If either $\mu(A|B) = 0$ or $\mu(C|D) = 1$, (2) is trivial. If not, $\max \underline{G}|B^* = \underline{G}(\omega_1) = \mu(C|D) - \mu(A|B) \geq 0$, i.e. (2) holds. A similar line of reasoning applies to upper probabilities. □

Example 2. If $A \Rightarrow B_1 \Rightarrow B_0$, it is easy to check that $A|B_0 \leq_{GN} A|B_1$. This is in fact a special case of Example 1. By Proposition 1,

$$\mu(A|B_0) \leq \mu(A|B_1) \tag{6}$$

if μ is either a probability (well known from the product rule $P(A|B_0) = P(A|B_1)P(B_1|B_0)$) or an upper/lower probability which is W-coherent (established in a different way in [10], Proposition 13) or C-convex. If $B_1 \Rightarrow B_0$, the GN relation and hence (6) still hold if $A \Rightarrow B_1 \vee B_0^c$, while, when A is arbitrary, (6) is replaced by

$$\mu(A \wedge B_1|B_0) \leq \mu(A|B_1). \tag{7}$$

To see this, use (6) and Proposition 1: $A \wedge B_1|B_0 \leq_{GN} A \wedge B_1|B_1 = A|B_1$.

3.1 The GN Relation in Extension Problems

An interesting feature of the GN relation is that it allows comparing conditional events whose conditioning events are possibly different. This fact is useful in the generalisations of the extension problem presented in the introduction that we are going to discuss now.

Let $\mathbb{P}$ be any partition. We wish to extend an uncertainty measure μ, assessed on $\mathcal{A}_C(\mathbb{P})$, to an arbitrary event $C|D$. We assume in what follows $C|D \neq \varnothing|D$, $C|D \neq D|D$, ruling out limiting cases where the extension is already known by (3).

Definition 4. Define $m(C|D) = \{A|B \in \mathcal{A}_C(\mathbb{P}) : A|B \leq_{GN} C|D\}$, $M(C|D) = \{A|B \in \mathcal{A}_C(\mathbb{P}) : C|D \leq_{GN} A|B\}$.

It is easy to see that

Proposition 2. *The sets m, M are non-empty and have, respectively, a maximum $(C|D)_*$ and a minimum $(C|D)^*$ conditional event w.r.t. $\leq_{GN}$,*

$$\begin{aligned} (C|D)_* &= (C \wedge D)_*|[(C \wedge D)_* \vee (C^c \wedge D)^*], \\ (C|D)^* &= (C \wedge D)^*|[(C \wedge D)^* \vee (C^c \wedge D)_*]. \end{aligned} \tag{8}$$

where $(C \wedge D)_* = \vee\{e \in \mathbb{P} : e \Rightarrow C \wedge D\}$, $(C^c \wedge D)^* = \vee\{e \in \mathbb{P} : e \wedge C^c \wedge D \neq \varnothing\}$, $(C \wedge D)^* = \vee\{e \in \mathbb{P} : e \wedge C \wedge D \neq \varnothing\}$, $(C^c \wedge D)_* = \vee\{e \in \mathbb{P} : e \Rightarrow C^c \wedge D\}$.

It holds that

Proposition 3. *Let $P(\cdot|\cdot)$ be a coherent precise probability on $\mathcal{A}_C$. Any of its extensions on $\mathcal{A}_C \cup \{C|D\}$ is a coherent precise probability iff $P(C|D) \in [P((C|D)_*); P((C|D)^*)]$.*

For extensions on an *arbitrary* set of conditional events, we have:

Proposition 4. *Let $\underline{P}$ ($\overline{P}$) be a W-coherent, respectively C-convex lower (upper) probability defined on $\mathcal{A}_C$ and $\mathcal{E}$ be an arbitrary set of conditional events. Then, the extension of $\underline{P}$ ($\overline{P}$) on $\mathcal{A}_C \cup \mathcal{E}$, such that $\underline{P}(C|D) = \underline{P}((C|D)_*)$ ($\overline{P}(C|D) = \overline{P}((C|D)^*)$), $\forall C|D \in \mathcal{E}$, is a W-coherent, respectively C-convex lower (upper) probability.*

In the special case of a coherent (precise) probability P, Propositions 3 and 4 show that both its extension $P(C|D) = P((C|D)_*)$, $\forall C|D \in \mathcal{E}$, and its extension $P(C|D) = P((C|D)^*)$, $\forall C|D \in \mathcal{E}$, are surely again coherent precise probabilities on $\mathcal{A}_C \cup \mathcal{E}$ only when $\mathcal{E}$ is a singleton. In general, they are either a lower (the former) or an upper (the latter) W-coherent probability on $\mathcal{A}_C \cup \mathcal{E}$. Like its unconditional counterpart, this extension problem naturally generates imprecise uncertainty measures. When the initial measure is a lower (upper) W-coherent probability $\underline{P}$ ($\overline{P}$), so is its extension by Proposition 4. Let us term it *GN-extension*. Interestingly, as a straightforward consequence of Propositions 1 and 2, the GN-extension is the *least committal* extension of $\underline{P}$ (or $\overline{P}$): in fact, taking for instance $\underline{P}$, for any other W-coherent extension $\underline{Q}$ of $\underline{P}$, $\underline{Q}(C|D) \geq \underline{Q}((C|D)_*) = \underline{P}((C|D)_*) = \underline{P}(C|D)$ (for the first equality, note that $\underline{Q}$ and $\underline{P}$ coincide on $\mathcal{A}_C$). As well known, this property identifies the *natural extension* (cf. [11]) of $\underline{P}$ on $\mathcal{A}_C(\mathbb{P}) \cup \mathcal{E}$. Similarly, when the initial $\underline{P}$ (or $\overline{P}$) is C-convex the GN-extension is what is termed *C-convex natural extension* in [7]. In both instances, computing the natural extensions is straightforward using the GN-relation: we just have to detect $(C|D)_*$ (or $(C|D)^*$).

The procedure requires that the starting P, $\underline{P}$ or $\overline{P}$ are defined on $\mathcal{A}_C$. If they are assessed on an arbitrary set S of conditional events, we should first extend P, $\underline{P}$ or $\overline{P}$ on some $\mathcal{A}_C(\mathbb{P}) \supset S$ to apply Proposition 4; a convenient $\mathbb{P}$ is the partition generated by S. Clearly, Proposition 4 does not add much, *operationally*, in this case: we would still need an operational procedure for the extension on $\mathcal{A}_C(\mathbb{P})$. It is meaningful at a *theoretical* level, as an explanation of how logical constraints determine our inferences on additional events.

4 A GN Type Relation with Imprecise Previsions

How could the GN relation $\leq_{GN}$ be defined and employed to compare conditional random numbers? We next propose a possible generalisation.

Definition 5. $X|B \leq_{GN} Y|D$ iff

$$BX(\omega) + \sup_B X \cdot B^c D(\omega) \leq DY(\omega) + \inf_D Y \cdot BD^c(\omega), \forall \omega \Rightarrow B \vee D. \quad (9)$$

The motivation for this definition is very similar to the betting argument in [5, 6], recalled in Section 2 to justify the GN relation. In fact:

a) if $\omega \Rightarrow B \wedge D$, (9) reduces to $X(\omega) \leq Y(\omega)$. This means: whenever we bet both on $X|B$ and on $Y|D$, we gain at least as much with the bet on $Y|D$;
b) for $\omega \Rightarrow B^c \wedge D$, it reduces to $\sup_B X = \sup\{X|B\} \leq Y(\omega)$. For such ω, we bet on $Y|D$ but not on $X|B$. By the last inequality the gain from our bet on $Y|D$ is not less than our (potential) gain on $X|B$, had we bet on it;
c) if $\omega \Rightarrow B \wedge D^c$, (9) reduces to $X(\omega) \leq \inf_D Y = \inf\{Y|D\}$, whose interpretation is specular to that in b) above.

When $X|B \neq \varnothing|B$ and $Y|D \neq D|D$, Definition 5 generalises Definition 3: if X, Y are (indicators of) events, say $X = A$, $Y = C$, (9) becomes $AB + \max\{A|B\} \cdot B^c D \leq CD + \min\{C|D\} \cdot BD^c$, and it can be shown that $A|B \leq_{GN} C|D$ (by Definition 3) iff this inequality holds.

The partial ordering *among conditional random numbers* of the generalised GN relation induces an agreeing ordering on their uncertainty measures in the cases stated by the next result:

Proposition 5. *Let* $S = \{X|B, Y|D\}$. *Then*

$$X|B \leq_{GN} Y|D \Rightarrow \mu(X|B) \leq \mu(Y|D) \quad (10)$$

whenever μ *is either a coherent precise prevision* P *or a W-coherent lower (upper) prevision* $\underline{P}$ *(*$\overline{P}$*), defined on* S.

Proposition 5 generalises Proposition 1, except for C-convex probabilities. It is not clear at present whether the result applies to such previsions too.

As for the inferential use of the GN relation with conditional random numbers, the problem is considerably more complex than with events and largely to be investigated yet. Consider for this the following example:

Example 3. How does (7) in Example 2 generalise with random numbers? Intuitively, we should first check whether, replacing A with X, $B_1X|B_0 \leq_{GN} X|B_1$ is still true, if $B_1 \Rightarrow B_0$. Ignoring the trivial case $B_0 = B_1$, by Definition 5 this inequality is equivalent to $B_1B_0X + \sup_{B_0} X \cdot (B_0^c B_1) \leq B_1X + \inf_{B_1} X \cdot (B_0B_1^c)$, which is equivalent ($B_1B_0 = B_1$, $B_0^c B_1 = 0$, $B_0B_1^c \geq 0$) to $\inf(X|B_1) \geq 0$, a condition always fulfilled if $X|B$ is a conditional event. On the contrary, it is easy to check that $X|B_1 \leq_{GN} B_1X|B_0$ iff $\sup(X|B_1) \leq 0$. Hence, $X|B_1$ and $B_1X|B_0$ are GN-incomparable iff $\inf(X|B_1) \cdot \sup(X|B_1) < 0$. In this example, the GN relation is *sign sensitive*.

5 Conclusions

With imprecise probabilities, the GN relation allows computing easily the natural extension of a W-coherent (or C-convex) assessment on the special set $\mathcal{A}_C$, a problem which may be viewed as a generalisation of de Finetti's Fundamental Theorem. While the exact relevance of the GN relation with imprecise previsions remains an open question, it may be asserted that it supplies bounds on uncertainty measures, in certain specific instances. One such case is considered in Example 3. However, the bound it achieves is rather rough. For a stricter bound under looser hypotheses, see Proposition 3.1 in [9]. For another example, let $X|B \geq 0$ take up a finite number of values $x_1, x_2, \ldots, x_n$. Then W-coherence requires that $\underline{P}(X|B) = \underline{P}(\sum_{i=1}^{n} x_i e_i|B) \geq \sum_{i=1}^{n} x_i \underline{P}(e_i|B)$, where e_i is the event '$X = x_i$', $i = 1, \ldots, n$. Applying Proposition 1 to each $\underline{P}(e_i|B)$, we get the lower bound $\underline{P}(X|B) \geq \sum_{i=1}^{n} x_i \underline{P}((e_i|B)_*)$.

References

1. Coletti, G., Gilio, A., Scozzafava, R.: Comparative probability for conditional events: A new look through coherence. Theory and Decision 35(3), 237–258 (1993)
2. Coletti, G., Scozzafava, R.: Characterization of coherent conditional probabilities as a tool for their assessments and extensions. Int. J. Uncertain. Fuzz. 4(2), 103–127 (1993)
3. de Finetti, B.: Theory of Probability. John Wiley & Sons, Ltd., New York City (1974)
4. Goodman, I.R., Nguyen, H.T.: Conditional objects and the modeling of uncertainties. In: Gupta, M., Yamakawa, T. (eds.) Fuzzy Computing, pp. 119–138. North Holland, Amsterdam (1988)
5. Milne, P.: Bruno de Finetti and the logic of conditional events. Br. J. Philos. Sci. 48(2), 195–232 (1997)
6. Milne, P.: Bets and boundaries: assigning probabilities to imprecisely specified events. Studia Logica 90, 425–453 (2008)
7. Pelessoni, R., Vicig, P.: Uncertainty modelling and conditioning with convex imprecise previsions. Int. J. Approx. Reason. 39(2-3), 297–319 (2005)
8. Pelessoni, R., Vicig, P.: Williams coherence and beyond. Int. J. Approx. Reason. 50(4), 612–626 (2009)
9. Pelessoni, R., Vicig, P.: Bayes' Theorem Bounds for Convex Lower Previsions. J. Stat. Theor. Pract. 3(1), 85–101 (2009)
10. Pelessoni, R., Vicig, P., Zaffalon, M.: Inference and risk measurement with the parimutuel model. Int. J. Approx. Reason. 51(9), 1145–1158 (2010)
11. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London (1991)
12. Williams, P.M.: Notes on conditional previsions. Int. J. Approx. Reason. 44(3), 366–383 (2007)

The CONEstrip Algorithm

Erik Quaeghebeur

Abstract. Uncertainty models such as sets of desirable gambles and (conditional) lower previsions can be represented as convex cones. Checking the consistency of and drawing inferences from such models requires solving feasibility and optimization problems. We consider finitely generated such models. For closed cones, we can use linear programming; for conditional lower prevision-based cones, there is an efficient algorithm using an iteration of linear programs. We present an efficient algorithm for general cones that also uses an iteration of linear programs.

Keywords: Consistency, convex cones, feasibility, inference, linear programming.

1 Introduction

Mathematically speaking, frameworks for modeling uncertainty consist of rules that specify what constitutes a within the framework valid model and rules to perform computations with such models. For a number of frameworks under the imprecise probability umbrella [8, 9], checking validity—i.e., the consistency criteria of avoiding sure & partial loss and coherence—and calculating an inference—i.e., natural extension—involves solving feasibility and optimization problems.

We illustrate in Section 2 that the feasibility aspect of these problems essentially boils down to checking whether some vector lies in a *general convex cone*, called *general cone* from now on, a cone that may be closed, open, or *ajar*, i.e., neither open nor closed. For models specified in a finitary way, algorithms to do this for closed and specific general cases can be found in

Erik Quaeghebeur
SYSTeMS Research Group, Ghent University, Belgium
e-mail: erik.quaeghebeur@ugent.be

R. Kruse et al. (Eds.): Synergies of Soft Computing and Statistics, AISC 190, pp. 45–54.
springerlink.com