Seminario POLIS di Inverno - 24 Gennaio 2014

Some measures to evaluate the importance of factors contributing in a process

Vito Fragnelli

Università del Piemonte Orientale

vito.fragnelli@mfn.unipmn.it

Joint work with:

Josep Freixas - josep.freixas@upc.edu
Montserrat Pons - montserrat.pons@upc.edu
Lluís Sanmiquel - sanmi@emrn.upc.edu

1984

Summary

Motivation

A Preliminary Example

Theoretical Issues

Relevance Measures

Properties

Concluding Remarks

Motivation

Game theory studies conflicts and cooperation between rational decision-makers, but it is possible to apply some of its techniques when players are not rational decision-makers
Examples can be found in medicine (Fragnelli and Moretti, 2008, Lucchetti, Radrizzani, and Munarini, 2011, Moretti, Fragnelli, Patrone, and Bonassi, 2010) or in engineering (Aven and Østebø, 1986, Boland and El-Neweihi, 1995, Fragnelli, García-Jurado, Norde, Patrone and Tijs, 1999, Freixas and Pons, 2008, Kuo and Zhuo, 2012)

The Shapley value (Shapley, 1953) is a classical way to allocate the profit of cooperation
Moretti and Patrone (2008) survey the transversal use of this value, with applications to medicine, reliability and telecommunications, in which the players are genes, components and antennas, respectively
Lipovetsky and Conklin (2001) apply the Shapley value to multiple regression analysis for estimating the relative impact of the different predictors

HERE

A different and novel use of cooperative games for obtaining a direct measure of the importance of the factors contributing in a process

Factors are not necessarily rational decision-makers, so the model is rooted in the notion of *incidence function* instead of that of characteristic function of the cooperative game, and the aims are *relevance measures* for factors instead of values for players

The proposed methodology may be applied to the analysis of factors in traffic accidents, in mining accidents, quality control analysis, diseases, etc.

The information available is the data obtained from the different repetitions of the process in the period under analysis, and a set of pre-selected factors that may influence the process

These data lead to an *incidence function* a that associates to each subset of factors the number of times that them (and only them) are present in the process

Assumptions

- no other information on the process, apart of the incidence function a, is available
- the set of pre-selected factors is clearly known
- the factors are mutually independent

Given an incidence function a, a *relevance measure* assigns a real number to each factor, stressing its level of importance

The relevance measures can be used for different purposes such as:

- budget distribution for improving the future occurrences of the process
- checking if previous policies were effective
- performing subsequent complementary studies when some exogenous information not encapsulated in the incidence function a is available

A Preliminary Example

Periodically, a small university department shares a fraction of its resources among the members that have done scientific research in a fixed previous period of time according to their paper authorship in specialized journals
N is the set of researchers of the department, $a(S), S \subseteq N$ is the number of papers coauthored by all the members of S published within the given period; coauthorship with outsiders is not taken into account
The department has only four members, three senior researchers r, s and t, and a young researcher y, i.e. $N = \{r, s, t, y\}$

Consider the following scenarios:

1. Each published paper is rewarded equally and the spoil is equally divided among its authors
2. Coauthorship within members of the department is stimulated and the spoil per paper is divided among authors equally
3. Publication is encouraged for the young researcher, no matter the number of coauthors for her publications, and coauthorship with the young researcher is stimulated for the senior researchers

The incidence function is:

$$\begin{array}{llll} a(r) = 10, & a(s) = 6, & a(t) = 2, & \\ a(rt) = 2, & a(st) = 2, & a(sy) = 1, & a(ty) = 3, \\ a(sty) = 3, & & & \\ a(S) = 0 & otherwise & & \end{array}$$

The researchers of the department published 29 papers

r is by far the more productive and is the sole author of most of his publications

t is considerably less active in publishing alone, but she is the most cooperative senior researcher

s plays an intermediate role

- Scenario 1

Each published paper is scored 1, that is equally divided among the authors

The resulting measure is:

$$\mathcal{F}_i(a) = \sum_{S \subseteq N : i \in S} \frac{a(S)}{|S|}, \quad i \in N$$

that for the example gives:

$$\mathcal{F}(a) = (11, 8.5, 6.5, 3)$$

The budget would be divided proportionally to these weights

- Scenario 2

Incentives are provided for coauthored papers

The score of each paper is still equally divided among the authors, but the score assigned to joint papers linearly increases with the number of coauthors: an article with a single author is scored 1 and an extra score of ε is added for each additional author after the first one

The measure is:

$$\mathcal{S}_i(a) = \sum_{S \subseteq N : i \in S} \frac{a(S)(1 + \varepsilon(|S| - 1))}{|S|}, \quad i \in N$$

Applying it to the example with $\varepsilon = 0.5$, the weights are 1, 1.5, 2, and 2.5 for papers of 1, 2, 3, and 4 authors, respectively and the measure gives:

$$\mathcal{S}(a) = (11.5, 10.25, 9.25, 5)$$

Setting $\varepsilon = 1$, the measure gives:

$$\tilde{\mathcal{S}}(a) = (12, 12, 12, 7)$$

i.e. the number of papers published by each author

It is possible to provide stronger incentives for collaboration only among two researchers, scoring 2 the papers with 2 authors and 1 the others; the measure results to be:

$$\bar{\mathbb{S}}_i(a) = \sum_{S \subseteq N : i \in S, |S| \neq 2} \frac{a(S)}{|S|} + \sum_{S \subseteq N : i \in S, |S| = 2} a(S), \quad i \in N$$

which, applied to the example gives:

$$\bar{\mathbb{S}}(a) = (12, 10, 13, 5)$$

- Scenario 3

It is possible not to treat symmetrically the seniors and the young researchers

All the publications by the young researcher are scored 1 for her. For senior researchers the weight is $1/(|S|-1)$ if the young researcher is coauthor and $1/|S|$ otherwise

The measure is defined as:

$$\mathcal{P}_i(a) = \begin{cases} \sum_{S\subseteq N: i\in S, y\notin S} \dfrac{a(S)}{|S|} + \sum_{S\subseteq N: i,y\in S} \dfrac{a(S)}{|S|-1}, & \text{if } i \text{ is a senior researcher} \\ \\ \sum_{S\subseteq N: i\in S} a(S), & \text{if } i \text{ is the young researcher} \end{cases}, \; i\in N$$

that for the example gives:

$$\mathcal{P}(a) = (11, 9.5, 8.5, 7)$$

Comparison of the measures:

	r	s	t	y
$\mathcal{F}/29$	0.379	0.293	0.224	0.103
$\mathcal{S}/36$	0.319	0.285	0.257	0.139
$\tilde{\mathcal{S}}/43$	0.279	0.279	0.279	0.163
$\bar{\mathcal{S}}/40$	0.300	0.250	0.325	0.125
$\mathcal{P}/36$	0.306	0.264	0.236	0.194

Table 1: Normalized measures

$\mathcal{F}$ penalizes the young researcher y, who is favored by $\mathcal{P}$. On the other hand, $\mathcal{F}$ is the best option for r and s, while t would prefer $\bar{\mathcal{S}}$. Finally, s could be quite indifferent among the five measures

Theoretical Issues

Let $\mathcal{P}$ be a process and $N = \{1, 2, \ldots, n\}$ be the selected set of significant independent factors intervening in it
An incidence function on N is a function $a : 2^N \to \mathbb{R}_{\geq}$ such that $a(\emptyset) = 0$; a assigns to any subset S of N ($S \neq \emptyset$) the number of occurrences of $\mathcal{P}$ in which all the factors in S intervened, but none of the factors in $N \setminus S$
The function a by its nature fails to fulfill some properties promoting cooperation among decision-makers as monotonicity, convexity or superadditivity
Let $\mathcal{A}^N$ be the class of all incidence functions on N; it is possible to define two natural operations on $\mathcal{A}^N$, the *sum* and the *product for a non-negative real number*, which give new incidence functions:

- If a_1, $a_2 \in \mathcal{A}^N$: $(a_1 + a_2)(S) = a_1(S) + a_2(S)$ for every set of factors $S \subseteq N$
- If $a \in \mathcal{A}^N$ and $k \in \mathbb{R}_{\geq}$: $(ka)(S) = k \cdot a(S)$ for every set of factors $S \subseteq N$

$\mathcal{A}^N$ assumes the structure of a cone in $\mathbb{R}^{2^N-1}$ with the null incidence function η defined by $\eta(S) = 0$ for all set of factors $S \subseteq N$ as proper zero element in $\mathcal{A}^N$
Let $T(a) = \sum_{S \subseteq N} a(S)$ be the total number of occurrences of $\mathcal{P}$

Relevance Measures

Definition 1 (Relevance measure) *A* relevance measure *is a function* $f : \mathcal{A}^N \longrightarrow \mathbb{R}^N_{\geq}$ *that assigns to every incidence function,* a*, the vector* $(f_1(a), f_2(a), \ldots, f_n(a))$ *where the non-negative real number* $f_i(a), i \in N$ *is interpreted as the importance of factor* i *in the process associated to the incidence function* a

Different relevance measures can be defined on an incidence function $a \in \mathcal{A}^N$; let $i \in N$ be a generic factor

The egalitarian measure $\mathfrak{e}$

$$\mathfrak{e}_i(a) = T(a)/n$$

It assigns the same value to all factors, independently of the frequency in which they appear

It is a solidarity measure

The basic measure $\mathfrak{b}$

$$\mathfrak{b}_i(a) = \sum_{S \subseteq N \,:\, i \in S} a(S)$$

It is the second one proposed in scenario 2; it seems very natural when any factor is able to generate the outcome even independently from the others

The fair measure $\mathfrak{F}$

$$\mathfrak{F}_i(a) = \sum_{S\subseteq N\colon i\in S} \frac{a(S)}{|S|}$$

It is the one proposed in scenario 1; it is the natural measure to be chosen if all factors in each set are supposed to have the same *a priori* weight and each occurrence of the process is treated equally

The weighted measures $\mathfrak{b}^c$

$$\mathfrak{b}_i^c(a) = \sum_{S\subseteq N\colon i\in S} a(S)c(i,S)$$

where $c : N \times 2^N \longrightarrow \mathbb{R}$ is a function which allows to weight subsets in a different way for any $i \in N$

The basic and the fair measures are particular cases of these measures when $c(i,S) = 1$ and $c(i,S) = \dfrac{1}{|S|}$, respectively, for any $i \in N$; all the measures in the preliminary example are of this kind

The selective measures $\mathfrak{s}^\alpha$

$$\mathfrak{s}_i^\alpha(a) = \sum_{S \subseteq N \,:\, i=\alpha(S)} a(S)$$

where α is a selection function, $\alpha : 2^N \to N$, with $\alpha(S) \in S$ for all $S \neq \emptyset$ It seems very natural when for each set of factors S it is possible to assign the whole importance to factor $\alpha(S)$, i.e. the other factors in S, if any, depend on $\alpha(S)$

The proportional measure $\mathfrak{p}$

$$\mathfrak{p}_i(a) = \frac{T(a)}{\sum\limits_{j \in N} a(\{j\})} \cdot a(\{i\})$$

It is well-defined if in at least one performance of the process only a single factor occurred; it seems very natural when it is not sure that when a performance involves more than one factor all the factors are really effective. Consider a road accident that involves a driver with serious damages on a car in bad condition, so it is difficult to say if these negative elements where already present before the accident, or one of the two is simply a consequence of the accident

Properties

Definition 2 *Let $a \in \mathcal{A}^N$*

- *A factor i is* null *in a if $a(S) = 0$ for all $S \subseteq N$ with $i \in S$*
- *Two different factors i and j have* equivalent incidence *in a if $a(S \cup \{i\}) = a(S \cup \{j\})$ for all $S \subseteq N \setminus \{i, j\}$*

Definition 3 *A relevance measures f satisfies the property of:*

- Totality: *if $\sum_{i \in N} f_i(a) = T(a)$ for all $a \in \mathcal{A}^N$*
- Zero on nulls: *if $f_i(a) = 0$ for any factor i null in $a \in \mathcal{A}^N$*
- Equal treatment: *if $f_i(a) = f_j(a)$ for all pairs of factors i and j with equivalent incidence in $a \in \mathcal{A}^N$*
- Linearity: *if $f_i(\alpha a + \beta b) = \alpha f_i(a) + \beta f_i(b)$ for all $\alpha, \beta \in \mathbb{R}_{\geq}, a, b \in \mathcal{A}^N$ and for all $i \in N$*
- Monotonicity: *if $a(S) \geq b(S)$ for all $S \subseteq N, S \ni i$ implies $f_i(a) \geq f_i(b)$ for all $a, b \in \mathcal{A}^N$*

Linearity allows for weighted combinations of different incidence functions

Monotonicity tells that if a factor i has larger incidence in function a than in function b, then the relevance for factor i in function b should be at most the same as in function a

	Totality	Zero on nulls	Equal treatment	Linearity	Monotonicity
Egalitarian	Yes	No	Yes	Yes	No
Basic	No	Yes	Yes	Yes	Yes
Fair	Yes	Yes	Yes	Yes	Yes
Weighted	No*	Yes	No*	Yes	Yes
Selective	Yes	Yes	No	Yes	Yes
Proportional	Yes	Yes	Yes	No	No

Table 2: Properties

No* means that the property is verified or not depending on the considered weights

Notice that the fair relevance measure is the unique of the former measures which verifies all of these properties

Proposition 1 *Let f be a relevance measure.*

- *If f verifies Totality, Monotonicity and Equal treatment properties then it also verifies Zero on nulls property*
- *If f verifies Totality, Monotonicity and Equal treatment properties then it also verifies Linearity property*

Proposition 2 *There exists only one relevance measure that satisfies Totality, Zero on nulls, Equal treatment and Linearity properties. This measure is precisely the fair relevance measure*

Example 1 (Independence)

- *The basic measure $\mathfrak{b}$ satisfies Zero on nulls, Equal treatment and Linearity but it does not verify Totality*
- *The egalitarian measure $\mathfrak{e}$ satisfies Totality, Equal treatment and Linearity, but it does not satisfy Zero on nulls*
- *The selective measure $\mathfrak{s}^{\alpha}$ satisfies Totality, Zero on nulls and Linearity, but it does not satisfy Equal treatment*
- *The proportional measure $\mathfrak{p}$ satisfies Totality, Zero on nulls and Equal treatment, but it does not satisfy Linearity*

Theorem 1 *There exists only one relevance measure that satisfies Totality, Equal treatment and Monotonicity properties. This measure is precisely the fair relevance measure*

This theorem is an immediate consequence of Propositions 1 and 2

Example 2 (Independence)

- *The basic measure* $\mathfrak{b}$ *satisfies Equal treatment and Monotonicity, but it does not verify Totality*
- *The selective measure* $\mathfrak{s}^\alpha$ *satisfies Totality and Monotonicity but it does not satisfy Equal treatment*
- *The proportional measure* $\mathfrak{p}$ *satisfies Totality and Equal treatment but it does not satisfy Monotonicity*

Concluding Remarks

- The factors were supposed to be independent, but sometimes this hypothesis may result too strong; it is possible that factors considered independent are connected in practice
- It is possible to refer to situations in which several factors are identified, so we can study the possibility of using approximated measures
- In case of a high number of factors, the reliability of the data may be questionable. In these cases it may be useful to use a subset of the available data, e.g. only those related to occurrences that are caused by at most two factors

Main References

T. Aven and R. Østebø (1986) Two new component importance measures for a flow network system. *Reliability Engineering*, 14:75–80

P.J. Boland and E. El-Neweihi (1995) Measures of component importance in reliability theory. *Computers Ops Res*, 4:455–463

V.Fragnelli, I.García-Jurado, H.Norde, F.Patrone and S.Tijs (1999) How to Share Railway Infrastructure Costs? In F.Patrone, I.García-Jurado, S.Tijs (eds.) *Game Practice: Contributions from Applied Game Theory*, 91–101, Kluwer, Amsterdam (NL)

V. Fragnelli and S. Moretti (2008) A game theoretical approach to the classification problem in gene expression data analysis. *Computers and Mathematics with Applications*, 55:950–959

J. Freixas and M. Pons (2008) Identifying optimal components in a reliability system. *IEEE Transactions on Reliability*, 57:163–170

W. Kuo and X. Zhuo (2012) *Importance measures in reliability, risk, and optimization.* Wiley

S. Lipovetsky and M. Conklin (2001) Analysis of regression in game theory approach. *Applied stochastic models in business and industry*, 17:319–330

R. Lucchetti, P. Radrizzani, and E. Munarini (2011) A new family of regular semivalues and applications. *International Journal of Game Theory*, 40:655–675

S. Moretti, V. Fragnelli, F. Patrone, and S. Bonassi (2010) Using coalitional games on biological networks to measure centrality and power of genes. *Bioinformatics*, 26:2721–2730

S. Moretti and F. Patrone (2008) Transversality of the Shapley value. *TOP*, 16:1–41

L.S. Shapley (1953) A value for n-person games. In H.W. Kuhn and A.W. Tucker (eds.) *Contributions to the Theory of Games II*, 307–317, Princeton University Press, Princeton, USA

Thanks!