Towards Causal Modeling of Human Behavior

Matteo Campo[1], Anna Polychroniou[1], Hugues Salamin[1], Maurizio Filippone[1], and Alessandro Vinciarelli[1,2]

[1] University of Glasgow - Sir A.Williams Bldg. - G12 8QQ Glasgow (UK)
[2] Idiap Research Institute - CP 592 - 1920 Martigny (Switzerland)
firstname.lastname@glasgow.ac.uk

Abstract. This article proposes experiments on decision making based on the "Winter Survival Task", one of the scenarios most commonly applied in behavioral and psychological studies. The goal of the Task is to identify, out of a predefined list of 12 items, those that are most likely to increase the chances of survival after the crash of a plane in a polar area. In our experiments, 60 pairs of unacquainted individuals (120 subjects in total) negotiate a common choice of the items to be retained after that each subject has performed the task individually. The results of the negotiations are analyzed in causal terms and show that the choices made by the subjects individually act as a causal factor with respect to the outcome of the negotiation.

1 Introduction

In the last years, automatic analysis of human behavior has attracted a large deal of attention in the computing community (see [1,2] for extensive surveys). The efforts focused on two main directions, namely (i) the synthesis of human behavior - in particular when it comes to social and affective phenomena that make embodied conversational agents believable, and (ii) the automatic understanding of human communication dynamics, with particular attention to the prediction of behavioral outcomes and the inference of socially relevant information from nonverbal communication.

Current approaches tend to adopt a purely computational perspective, i.e. they do not try to understand the phenomena they synthesize or analyze, but simply to maximize performance metrics like the recognition rate (percentage of times an approach makes the correct prediction) or the Mean Opinion Score (average appreciation score assigned by users). Such a perspective is certainly effective, but the interdisciplinary collaboration with human sciences, inevitable when dealing with human behavior, shows that no technology can be effective in the field without understanding human-human and human-machine interactions in terms of causes and effects [3,4,5].

The statistical literature has studied extensively approaches aimed at learning cause-effect relationships from data (see [6] for an extensive survey). However, these approaches were largely neglected in the computing community, in part

B. Apolloni et al. (Eds.): *Neural Nets and Surroundings*, SIST 19, pp. 337–344.
DOI: 10.1007/978-3-642-35467-0_33

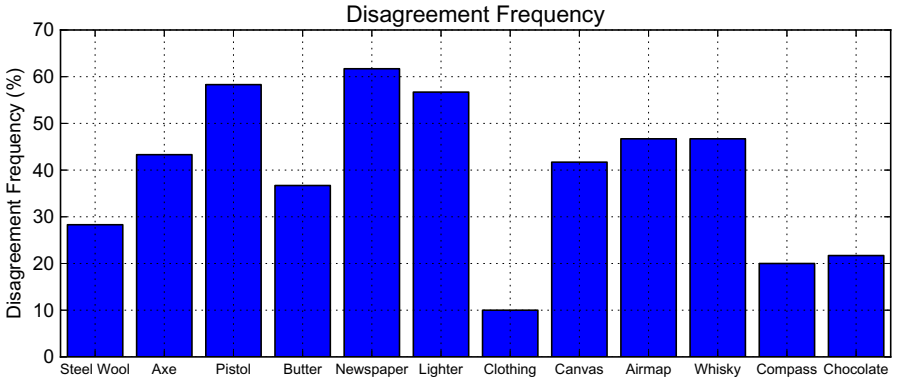

Fig. 1. The plot shows the percentage of calls where there was a disagreement on each item

because they require the formulation of untested causal assumptions about the phenomena under analysis, in part because they adopt notations and terminology different from those commonly used in the machine intelligence community [7]. This paper shows that overcoming these two barriers can be beneficial to the computational analysis of social interactions from at least two points of view. The first is that it makes social phenomena more predictable and, hence, easier to analyze automatically; the second is that it allows a better understanding of the data being modeled.

2 The Experiment

The experiment adopts the "*Winter Survival Task*", a scenario where participants must identify, out of a list of 12 predefined items, those that maximize the chances of survival after an emergency landing in Northern Canada (in the middle of the winter). The main advantage of the scenario, often used in psychological and behavioral experiments, is that the average subject is unlikely to have experienced a plane crash or to know survival techniques suitable for a winter beyond the Polar Circle. Hence, the outcomes of the experiment are likely to depend on social and psychological phenomena during the interactions and not on skills and knowledge the participants have before and independently of the experiment.

In this work, the participants earn three British Pounds each time they select a correct item (there is a gold standard for the task), but loose the same amount of money when they select a wrong one. In this way, whenever the subjects disagree about a certain item, they are motivated to persuade their interlocutor.

2.1 Experimental Protocol

The task was performed by 60 pairs of fully unacquainted subjects that have never met before the experiment (120 subjects in total). For each pair, the protocol included the following steps:

- The two subjects are accompanied to different rooms without meeting or crossing one another.
- Once in their room, the subjects receive a mobile phone (same model for all participants) and the documentation accompanying the experiment (scenario, questionnaires, etc.), including the list of the 12 items at the core of the Winter Survival Task.
- Before starting the call, the subjects fill a form where they must write a decision ("*Yes*" or "*No*") for each item of the list. This makes it possible to know, for each item, what is the decision made by each subject before any interaction with their counterpart.
- One of the two subjects, selected randomly, calls the other and starts the discussion item per item.
- During the call, the subjects discuss item by item and negotiate a common solution ("*Yes*" or "*No*") that is the final outcome of the task.

At the end of the call, it is possible to know what are the items on which the participants disagree and, most importantly, what are the subjects that persuade their interlocutors in case of disagreement, i.e. the subjects that convince others to adopt their initial decision in case of disagreement. Figure 1 shows, for each item, the percentage of calls where discussion was needed to reach a common decision. While some items (e.g., the clothing) were discussed only a few times, others were frequently debated between participants.

3 Causal Analysis

The main question behind the experiment is what are the causal factors that increase the chances of persuading others. In other words, whether the decision about a given item is random (which is what the scenario seems to suggest), it depends on the characteristics of the subjects, or on their choice prior to the discussion. The next sections show how the problem was modeled and the resulting findings.

3.1 Modeling

The problem can be modeled with a set $\mathbf{X}$ of observable binary variables:

Role $\mathbf{R}$:	$R \in \{0, 1\}$
Gender $\mathbf{G}$:	$G \in \{0, 1\}$
Initial Choice $\mathbf{Y}$:	$Y \in \{0, 1\}$
Result $\mathbf{W}$:	$W \in \{0, 1\}$

The variable R stands for the role a participant had in the conversation in terms of "*caller*", the subjects who makes the phone call, or "*receiver*", the subject who receives the call (associated to 1 and 0, respectively). The variable G corresponds to the gender of the participant, with 0 and 1 for male and female, respectively. The variable Y accounts for the initial decision of the participants, with 0 and 1 corresponding to "*No*" and "*Yes*", respectively. Finally, the variable W accounts for whether the subject persuades the other person ($W = 1$) or not ($W = 0$).

The aim of this work is to make statements about the causal relationship between Y and W. In particular, we would like to estimate the following quantities:

Post-intervention distribution: The post-intervention distribution estimates the probability of a subject persuading or not the other person given that the "treatment" $Y = 1$ is applied (i.e., given that the initial decision of the subject is "*Yes*"):

$$p(W|\text{do}(Y = 1))$$

where $\text{do}(Y = 1)$ is the "do" operator [8] and expresses the probability of an effect given an "action" on the model (the action will be, for our purposes, taking an initial choice Y). It is important to point out that this quantity has a different and stronger meaning than the probability of an effect given an "observation", which is the usual conditional probability.

Counterfactuals: The probabilities of a change of effect, given a change of treatment. These can be expressed in terms of the following distributions:

$$\text{PN} = \frac{p(W = 1) - p(W = 1|\text{do}(Y = 0))}{p(W = 1, Y = 1)}$$

$$\text{PS} = \frac{p(W = 1|\text{do}(Y = 1)) - p(W = 1)}{p(W = 0, Y = 0)}$$

$$\text{PNS} = p(W = 1|\text{do}(Y = 1)) - p(W = 1|\text{do}(Y = 0))$$

that account, respectively, for the probability of Y being a necessary cause of W, the probability of Y being a sufficient cause of W, and the probability of Y being a necessary and sufficient cause of W.

Within the framework of causal inference, causal dependencies between variables are encoded by means of Direct Acyclic Graphs (DAG) [8]. Therefore, DAGs represent assumptions on the causal relationships between variables that can then be inferred after data are observed. Once the DAG is built, the first step to undertake in order estimate causal effects and counterfactuals is to verify whether these can actually be computed: this problem is called "identifiability" [8]. At the core of identifiability in DAG is the d-separation criterion, by which it is possible to derive conditional independence relationships between variables. In our experiments, d-separation was checked using TETRAD, a publicly available software package dealing with causal models [9].

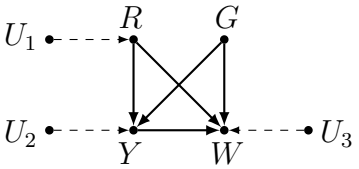

Fig. 2. The picture shows the Directed Acyclic Graph corresponding to the causal assumptions behind the experiments of this work

3.2 The Causal Model

This section shows the causal model used for the experiments of this work. The corresponding DAG is depicted in Figure 2 and the underlying causal assumptions are as follows:

- An unobserved variable U_1 influences the role of the participants;
- Role and gender influence the initial decision of the subjects and the result of the discussion;
- An unobserved variable U_2 influences the initial choice;
- The initial choice influences the result of the discussion;
- An unobserved variable U_3 influences the result of the discussion.

Unobserved variables are assumed to be deterministically related to their children and mutually independent. According to the graph (and its underlying assumptions), the joint probability distribution of the four observed variables is as follows:

$$p(\mathbf{X}) = p(R)\ p(G)\ p(Y|G,R)\ p(W|G,Y,R)\ . \tag{1}$$

Table 1. The table reports the Maximum Likelihood estimate of the joint probability distribution of $\mathbf{X}$ as obtained from the data observed in the 60 conversations used in the experiments

	$Y=0$	$Y=1$
$G=0$	0.036	0.021
$G=1$	0.026	0.016

$(R=0, W=0)$

	$Y=0$	$Y=1$
$G=0$	0.019	0.028
$G=1$	0.038	0.028

$(R=1, W=0)$

	$Y=0$	$Y=1$
$G=0$	0.071	0.123
$G=1$	0.075	0.134

$(R=0, W=1)$

	$Y=0$	$Y=1$
$G=0$	0.062	0.141
$G=1$	0.064	0.120

$(R=1, W=1)$

3.3 Post-Intervention Distribution

The model is *Markovian* because the associated graph is undirected and acyclic and the unobserved variables U_1, U_2, U_3 are mutually independent. This guarantees the causal effect $p(W|\text{do}(Y))$ to be identifiable. Performing an "intervention" on variable Y, that is, using the do-operator $\text{do}(Y = 1)$, the *post-intervention* distribution is:

$$p(W, R, G|\text{do}(Y = 1)) = p(R)\ p(G)\ p(W|Y = 1, R, G)\ .$$

Therefore, the following holds:

$$p(W = 1|\text{do}(Y = 1)) = \sum_{g,r\in\{0,1\}} p(r)\ p(g)\ p(W = 1|Y = 1, r, g) = 0.849$$

$$p(W = 1|\text{do}(Y = 0)) = \sum_{g,r\in\{0,1\}} p(r)\ p(g)\ p(W = 1|Y = 0, r, g) = 0.700\ .$$

The difference between the two probabilities is:

$$p(W = 1|\text{do}(Y = 1)) - p(W = 1|\text{do}(Y = 0)) = 0.149$$

and it expresses the difference between the effects of two different treatments. Note that the quantity above has also a counterfactual interpretation, as it is the probability of Y being a necessary and sufficient cause of W (PNS). A binomial test proportion with null hypothesis for $p(W = 1|\text{do}(Y = 1))$ and $p(W = 1|\text{do}(Y = 0))$ to be binomially distributed with the same success probability shows that the observed difference is statistically significant with p-value lower than $2 \cdot 10^{-4}$. This suggests that the value of Y acts as a causal factor and starting from an initial positive decision significantly increases the chances of persuading the counterpart.

3.4 Counterfactuals

The observed data show that the following relationship holds for the interactions used in the experiments:

$$p(W = 1|\text{do}(Y = 1)) \geq p(W = 1) \geq p(W = 1|\text{do}(Y = 0))$$

The relation above corresponds to the *monotonicity* of W relative to Y and it guarantees the identifiability of the three counterfactuals PN, PS, PNS [8]. According to the data, the values of the counterfactuals are:

$$\text{PN} = 0.171$$

$$\text{PS} = 0.503$$

The values are particularly interesting as they show the probability of obtaining a different result were the initial choices different. In particular, PN gives the

probability that changing the initial choice from $Y = 1$ to $Y = 0$ would have changed the result from $W = 1$ to $W = 0$. Conversely, PS gives the probability that changing the initial choice from $Y = 0$ to $Y = 1$ would have changed the result from $W = 0$ to $W = 1$. The latter represents the probability of Y being a sufficient cause of W, and suggests the interesting conclusion that in half of the cases where the initial choice $Y = 0$ led to $W = 0$, a different initial choice $Y = 1$ would have turn the result to $W = 1$.

4 Conclusions

This paper has presented a causal analysis of the decision-making behavior of individuals involved in the "Winter Survival Task". The experiments have involved 120 subjects and show that, when it comes to binary decisions about the acceptance of an item in the task, the initial choice of a subject acts as a causal factor for the final outcome of the discussion. In particular, subjects that start with an initial positive decision ("*the item should be retained*") have a probability of persuading others three times higher than the subjects starting with a negative decision ("*the item should not be retained*").

While being relatively simple, the experiments involve a large number of subjects that allow one to reliably estimate the causal effects. The main difference with respect to the application of traditional, associative statistics is that the estimated probabilities do not simply tell how frequently two or more variables take certain values, but what is the probability that one or more variables cause the value of one or more other variables. In this respect, the application promises to be fruitful not only from a technological point of view, making the prediction of interaction outcomes easier, but also from a scientific point of view, providing explanations about the observed results.

In the case of these experiments, all observations of interest could be modeled with binary variables, but in real-world scenarios, variables of interest are more likely to be continuous. This does not represent a major problem because all equations used in this work are independent of the actual expression used to estimate the probabilities. In other words, the tables used in these experiments can be replaced by distributions as complex as necessary without changing model assumptions, identifiability considerations, and formulas estimating the post-intervention probabilities or counterfactuals.

Future work will focus on those cases where causal modeling can actually make a major difference with respect to associative statistics, i.e. the analysis of non-experimental data. Those are data where conditions cannot be manipulated (as it typically happens in naturalistic settings for human-human and human-machine interactions), and hence, major effects might be missed by associative statistics simply because they are less frequent.

Acknowledgements. The Authors wish to acknowledge the support from the College of Science and Engineering of the University of Glasgow, the FP7 funded European Network of Excellence SSPNet, the Swiss National Center for Competence in Research on Interactive Multimodal Information Management (IM2) and the project "Human Emotional Interaction" funded by Finnish Ministry for the Technological Innovation (TEKES).

References

1. Vinciarelli, A., Pantic, M., Bourlard, H.: Social Signal Processing: Survey of an emerging domain. Image and Vision Computing Journal 27(12), 1743–1759 (2009)
2. Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., DÉrrico, F., Schroeder, M.: Bridging the gap between social animal and unsocial machine: A survey of social signal processing. IEEE Transactions on Affective Computing 3(1), 69–87 (2012)
3. Brunet, P., Cowie, R.: Towards a conceptual framework of research on social signal processing. Journal on Multimodal User Interfaces (to appear, 2012)
4. Mehu, M., Scherer, K.: A psycho-ethological approach to social signal processing. Cognitive Processing 13(2), 397–414 (2012)
5. Poggi, I., DÉrrico, F.: Social signals: a framework in terms of goals and beliefs. Cognitive Processing 13(2), 427–445 (2012)
6. Pearl, J.: Causal inference in statistics: An overview. Statistics Survey 3(1), 96–146 (2009)
7. Pearl, J.: Statistics and causal inference: A review. Test 12(2), 281–345 (2003)
8. Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press (2000)
9. Glymour, C., Scheines, R., Spirtes, P., Kelly, K.: Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling. Academic Press (1987)