Using KL Divergence for Credibility Assessment
Thibaut Vallée, Grégory Bonnet

To cite this version:

Thibaut Vallée, Grégory Bonnet. Using KL Divergence for Credibility Assessment. 2015 International Conference on Autonomous Agents and Multiagent Systems, May 2015, Istanbul, Turkey. hal-01196709

HAL Id: hal-01196709
https://hal.archives-ouvertes.fr/hal-01196709
Submitted on 10 Sep 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Using KL Divergence for Credibility Assessment

(Extended Abstract)

Thibaut Vallée, Grégory Bonnet
Normandie Univ, France
UNICAEN, GREYC, F-14032 Caen, France
firstname.lastname@unicaen.fr

ABSTRACT

In reputation systems, agents collectively estimate the others’ behaviours through feedbacks to decide with whom they can interact. To avoid manipulations, most reputation systems weight feedbacks with respect to the agents’ reputation. However, these systems are sensitive to some strategic manipulations, like oscillating attacks or whitewashing. In this paper, we propose (1) a credibility measure of feedbacks based on the Kullback-Leibler divergence to detect malicious behaviours and (2) filtering functions to enhance already known reputation functions.

Categories and Subject Descriptors
Computing methodologies [Artificial intelligence]: Distributed artificial intelligence

General Terms
Algorithms; Experimentation

Keywords
Trust and reputation; Collective intelligence

1. REPUTATION SYSTEM

In reputation systems, agents aggregate feedbacks from the others to evaluate the reliability of a given agent [2]. The robustness of such systems has been widely studied, and one of the most common manipulations is fake transaction or false feedback [1]. On the one hand, some recent works propose to split the agent’s reputation between its fiability as a service provider and its fiability (called credibility) as a feedback provider [3,4,6]. On the other hand, we drew an analogy between reputation systems and multi-armed bandit problems in our previous work [5]. In this article, we extend this previous work to feedbacks’ credibility assessment. We present firstly in Section 2 our model where agents collectively estimate through feedbacks the probability distributions that reward interactions. In order to detect false feedbacks and mitigate manipulations, we propose in Section 3 a credibility measure based on the Kullback-Leibler divergence between agent’s direct observations and feedbacks and filtering functions using this measure. First results against oscillating manipulations are given in Section 4.

2. INTERACTION MODEL

We consider a decentralized multi-agent system modeled by a multi-armed bandit where agents interact with each other in order to share services. Such system is modeled by a couple \((N,S)\) where \(N\) is the set of agents and \(S\) the set of available services. We denote by \(N_x \subset N\) the set of the agents that can provide the service \(s_x \in S\) and each agent \(a_k\) has an expertise factor \(\epsilon_{k,x} \in \mathbb{R}\) for each service it can provide. When an agent \(a_i\) asks a service \(s_x\) to an agent \(a_k\) at timestep \(t\), \(a_i\) receives a reward \(g_{i,k,x}^t \in \mathbb{R}\) that follows an unknown probability density function \(\Phi_{k,x}\) with \(\epsilon_{k,x}\) as its mean. The problem for an agent \(a_i\) needing a service \(s_x\) is to decide to which agent \(a_j \in N_x\) to ask this service in order to receive it with the best quality. We assume that the agents observe the rewards they received. An observation of \(a_i\) for the service \(s_x\) provided by \(a_k\) at the timestep \(t\) is thus the reward \(g^t_{i,k,x} \in \mathbb{R}\) and we denote by \(O_{i,k,x}\) the set of all observations received so far of \(a_i\) for the service \(s_x\) provided by \(a_k\). Even if the expertise of each agent is assumed to be unknown, the agents can estimate those values with their past observations. Hence, the agent \(a_i\) can use \(O_{i,k,x}\) to compute an estimation of \(\mu_{i,k,x}\) (mean) and \(\sigma_{i,k,x}\) (standard deviation) of \(\Phi_{k,x}\). In terms of reputation systems, the couple \((\mu_{i,k,x}, \sigma_{i,k,x})\) is \(a_i\)’s trust about \(a_k\)’s capability to provide the service \(s_x\). We assume then that each agent can receive feedbacks from other agents. Thus, agent \(a_i\) can provide to \(a_k\) a feedback about \(a_k\) and service \(s_x\) by communicating to \(a_i\) the set \(O_{j,k,x}\). We consider a reputation function \(f_i : N \times S \rightarrow \mathbb{R}\) which estimates according to agent \(a_i\), the expertise of an agent \(a_k\) in a given service based on a set of feedbacks. In the sequel, we consider two reputation functions: (1) individual estimation averages the own observations of agents \(a_i\); (2) collective estimation averages the own observations of agents \(a_i\) with all feedbacks it received. Notice that \(f_i\) may capture any other reputation function such as BetaReputation or EigenTrust.

3. CREDIBILITY MODEL

In order to mitigate manipulations, we propose a measure of feedback credibility to decide if they can be used in the reputation function. As feedbacks are an estimation of expected rewards, we propose to use the Kullback-Leibler divergence \(DKL\) as a credibility measure. Indeed, as the effective quality of services is assumed to follow an unknown probability density function, the divergence between
two honest feedbacks (that are a set of observations) tends towards 0 with a sufficient number of observations. Thus, if an agent provides a false feedback, the divergence should be greater than a given value. We set this value with the standard error of the mean (SEM) that represents the confidence of an agent for its own observations.

Definition 1. Let \( s_k \in S \) be a service and \( a_k \in N_x \) be an agent. Let \( a_i, a_j \in N \) be two agents. The divergence between the \( a_i \)'s observations and \( a_j \)'s feedbacks is:

\[
D_{i,j,k,x} = D_{KL}(N(\mu_{i,k,x}, \sigma_{i,k,x}^2) | N(\mu_{j,k,x}, \sigma_{j,k,x}^2))
\]

where \( N \) denotes the normal distribution.

Definition 2. Let \( s_k \in S \) be a service and \( a_k \in N_x \) be an agent. Let \( a_i, a_j \in N \) be two agents. A feedback \( O_{i,k,x} \) is KL-credible from \( a_i \)'s point-of-view\(^1\) if \( D_{i,j,k,x} \leq \delta \) where:

\[
\delta = D_{KL}(N(\mu_{i,k,x}, \sigma_{i,k,x}^2) | N(\mu_{i,k,x} + 1.96 \times \sigma_{i,k,x}, \sigma_{i,k,x}^2))
\]

We propose to enhance the reputation function \( f_i(a_k, s_k) \) (in this article the collective estimation) with a function that filter false feedbacks. We consider three filtering functions: (1) KL-filtering refuses all non KL-credible feedbacks; (2) \( k \)-fault filtering refuses all feedbacks from agents that provide at least \( k \) non KL-credible feedbacks; (3) \( k \)-lottocracy filtering accepts a feedback if and only if a random subset of \( k \) agents judge it KL-credible through majority voting.

4. EMPIRICAL STUDY

We evaluate the efficiency of our filtering functions against a coalition of malicious agents making an oscillating manipulation. This manipulation combines false feedbacks (both promotion and slandering) and whitewashing. Promotion aims at increasing the reputation of a given agent. Slandersing aims at decreasing an agent’s reputation. Whitewashing aims at resetting the malicious agent’s reputation in order to benefit a new reputation as a newcomer. We initialize the system with 100 agents including a coalition of 10 malicious agents. We compare the four reputation systems (collective estimation, KL-filtering, \( k \)-faults and \( k \)-lottocracy with \( k = 10 \)) with the individual estimation. Each simulation is repeated 50 times and we consider the mean regret. A small regret means that agents ask services to the best providers and, thus, are not influenced by false feedbacks. We ran three kinds of experiments: without and with malicious agents when the system is in nominal functioning and experiments with malicious agents when the system is initializing.

\(^1\)Here the 95% confidence interval of the SEM.

In nominal functioning, agents have already interacted and we compute only the regret of 20 new honest agents that join the system after 100 timesteps. The results are given in Figure 1. Obviously, when all agents are honest (Figure 1.1) sharing information decreases the regret. Most importantly, our filtering functions (except from \( k \)-lottocracy) provide the same regret than collective estimation. In presence of manipulations (Figure 1.2), collective estimation is clearly inefficient and our filtering functions allow to mitigate this problem (\( k \)-lottocracy is the most efficient in the first steps). However, in those results, newcomers can rely on agents that have a good knowledge of the system. In Figure 1.3, we consider that all agents are newcomers. As no agents have interacted, they firstly discard feedbacks and start with a high regret. After some timesteps, manipulations are detected as false feedbacks diverge from the agents’ observations (\( k \)-faults filtering is very efficient in this case).

5. CONCLUSION AND FUTURE WORKS

We modeled a multi-agent system where asking a service is equivalent to pulling an arm of a multi-armed bandit. Obviously, using a collective estimation decreases the agents’ regret but is sensitive to false feedbacks. We propose a feedback credibility measure based on Kullback-Leibler divergence and the agents’ confidence in their own observations. First results show that this is an efficient way to detect and discard false feedbacks. In future works, we need to evaluate our filtering functions on other reputation systems, such as BetaReputation or EigenTrust.

REFERENCES