Learning Initial Trust among Interacting Agents

Achim Rettinger[1], Matthias Nickles[1], and Volker Tresp[2]

[1] AI/Cognition Group, Technical University of Munich,
D-85748 Garching bei München, Germany,
{achim.rettinger, matthias.nickles}@cs.tum.edu,
[2] Corporate Technology, Siemens AG,
Information and Communications, Munich, Germany,
volker.tresp@siemens.com

Abstract. Trust learning is a crucial aspect of information exchange, negotiation, and any other kind of social interaction among autonomous agents in open systems. But most current probabilistic models for computational trust learning lack the ability to take context into account when trying to predict future behavior of interacting agents. Moreover, they are not able to transfer knowledge gained in a specific context to a related context. Humans, by contrast, have proven to be especially skilled in perceiving traits like trustworthiness in such so-called *initial trust situations*. The same restriction applies to most multiagent learning problems. In complex scenarios most algorithms do not scale well to large state-spaces and need numerous interactions to learn. We argue that trust related scenarios are best represented in a system of relations to capture semantic knowledge. Following recent work on nonparametric Bayesian models we propose a flexible and context sensitive way to model and learn multidimensional trust values which is particularly well suited to establish trust among strangers without prior relationship. To evaluate our approach we extend a multiagent framework by allowing agents to break an agreed interaction outcome retrospectively. The results suggest that the inherent ability to discover clusters and relationships between clusters that are best supported by the data allows to make predictions about future behavior of agents especially when initial trust is involved.

Keywords: Trust in Multiagent Systems, Information Agents, Agent Negotiation, Initial Trust, Relational Learning

1 Introduction

The assessment of trust values is getting increasingly important in distributed information systems since contemporary developments such as the Semantic Web, Service Oriented Architecture, Information Markets, Social Software, Pervasive and Ubiquitous Computing and Grid Computing are targeted mainly at open and dynamic systems with interacting autonomous entities. Such entities possibly show a highly contingent behavior, and it is often not feasible to implement effective mechanisms to enforce socially fair behavior as pursued in mechanism design or preference aggregation. Although computational trust has been focussed by research in Artificial Intelligence for several years (for an overview see [1]), current approaches still lack certain features of human trustability assessment which we consider to be of high importance for the computational determination of trust values in open systems. E.g., recent studies in psychology [2] have shown that people can robustly draw trait inferences like trustworthiness from the mere facial appearance of unknown people after a split second. Although seemingly neither the time span nor the available information allow to make a well-founded judgement, the derived trust (or distrust) provides after all a foundation for immediate decision making, and a significant reduction of social complexity especially under time pressure . Whereas the "quality" of such so-called *initial trust* (i.e., trusting someone without having accumulated enough experiences from relevant past behavior of the trustee) might be limited in the described scenario, this example shows that humans are able to estimate the trustability of others using information which are at a first glance unrelated to the derived expectation. (e.g., the facial appearance, or any contextual information in general). In contrast, the vast majority of approaches to empirical trust value learning in Artificial Intelligence lack this ability, as these approaches strongly rest on well-defined past experiences with the trustee, from which it is directly concluded that the trustee will behave in the future as he did in the past, regardless of the concrete context (cf. Section 6 for related work). These approaches come to their limits in cases where the trustor could not make such experiences and thus has to rely on "second order" information such as the context of the respective encounter instead. In order to make such initial trust computationally feasible, we not only need to relate trust values to a specific context, but we also need to provide a mechanism in order to take over contextualized trust to a new, possibly somewhat different context.

In particular, the general requirements that we wish to meet are:

Context sensitivity and trust transfer: Contextual information that might be related to the trust decision to be made needs to be incorporated. This shall include attributes of the person one needs to trust, attributes of the external circumstances under which the trust decision is made, and actions and promises the person has given to seek one's confidence. Furthermore, specific trust values gained in a certain context need to be *transferrable* to new, unknown "trigger" situations.

Multi-dimensionality: Most trust models assign a single trust value per agent. This ignores the fact that human trust decisions are made in relation to a whole spectrum of aspects (e.g., what a person is likely to do, such as the expected outcome of some information trading, even in the same context. For instance a certain information supplier agent might be trustworthy in terms of delivery date, but not in terms of information quality (e.g., precision, topicality, credibility...). Combining several trust related measures as in our approach is considerably much more flexible. In contrast, most existing approaches to trust still relate trust to "whole persons" only instead of their contextualized behavior.

At this, we focus on *interaction-trust* (i.e., (dis-)trust formed by agents during the course of an interaction regarding their opponents' behavior) in order to tailor our model to the specifics of the probably most relevant application field for empirical trustability assessment.

The remainder of this work is organized as follows: The next two Sections describes the basic scenario underlying our approach. Section 4 introduces our model for relational learning of initial trust, and Section 5 explores the general capabilities of our model with example data. Section 6 presents an application of initial trust learning in the context of simulated social interaction in order to provide a concrete evaluation of our approach. Section 7 discusses related work, and Section 8 outlines future research directions and concludes.

2 Modeling Interactions

Our scenario can be based on one of the most general frameworks for learning interactions in multiagent systems namely general-sum stochastic games (see [3]). A stochastic game can be represented as a tuple (A, C, Ac, R, T)[3]. A is the set of agents, C is the set of *stage games* (sometimes denoted as *states*), Ac is the set of actions available to each agent, R is the immediate reward function and T is a stochastic transition function, specifying the probability of the next stage game to be played.

It is in the nature of trust that we are dealing with incomplete and partially observable information. We neither assume the knowledge of the reward function R of the opponent nor their current state C. In fully observable games with perfect monitoring, incentives to betray can be estimated and trust becomes irrelevant because agents can be punished effectively [5]. Furthermore trust decisions require general sum games where joined gains can be exploited. Both zero-sum (e.g., [6]) and common-payoff (e.g., [7]) games are not relevant because either there are no joint gains or the agents' interests do not conflict.

Building on that formal setting our goal is to predict trust values O^e associated with the expectation of the next actions Ac given agent A and state C. We neither are trying to learn a strategy or policy nor are we interested in finding equilibria or proofing convergence. But we make contributions on how to scale MAL to more complex scenarios and show how an opponent model can be learned efficiently:

Predicting the next action of an opponent is an essential part of any model-based approaches to MAL [4]. The best-known instance of a model-based approach is fictitious play [9] where the opponent is assumed to be playing a stationary strategy. The opponent's past actions are observed, a mixed strategy is

[3] Our notation differs slightly from the commonly used ones, where A denotes actions and S states. Our notation should become clear in the next section

calculated according to the frequency of each action and then the best response is played, accordingly. This technique does not scale well to a large state-space $|C|$ as we experienced in our second experiment (Section 6): The same stage game is on average not observed before 400 interactions. Thus, this kind of naive approach does not allow to make an informed decision before 400 interactions and is obviously not suited for initial trust scenarios.

In our approach we make use of two techniques to face this issue. First, we allow to model any context related to the next trust-decision in a rich relational representation. This includes non-binding arrangements among agents also known as "cheap talk" [5] which take place before the actual interaction O^e is carried out and which are denoted as O^p. Second, we make use of techniques from the mature field of *Transfer Learning* [10] to reuse knowledge from previous interactions for potentially unknown future actions.

3 Modeling Interaction-Trust

The basic precondition for the emergence of trust are entities and social interactions between those entities. Hence, we chose a scenario that is interaction-centered as seen from the perspective of one agent who needs to trust (trustor) in someone/something (trustee). As usual in agent trust scenarios, (dis-)trust is related to the expected occurrence of some promised outcome (e.g., the communication of correct and precise information as negotiated before with some information trading agent, or the delivery of some other kind of product at the agreed price). The basic interaction-trust scenario then consists of:

1. A set of agents A (trustees) that are willing to interact with the trustor, each characterized by a set of observable attributes Att^A. An agent can be considered as a person or more general any instance that can be trusted, like an information source, a company, a brand, or an authority.
2. A set of external conditions or state C with corresponding attributes Att^C. An apparent condition would be the type of service provided by the trustee, for instance a specific merchandize or an information supply in case of information trading agents. Moreover this implies all external facts comprising this particular state like the trustor's own resources or the current market value of the merchandize in question.
3. A relation $interacts(a, c)$ with a set of relationship attributes Att^O capturing all negotiable interaction issues depending on a specific agent $a \in A$ and specific conditions $c \in C$. In general those attributes can be directly manipulated by the interacting agents and separated into two different sets:
   (a) Promised outcome O^p: Attributes Att^{O^p} of this set are (in general) observable before the trust-act is carried out.
   A typical attribute of this category is for example the price for the merchandize or the scope of the services offered, such as the amount and precision of information in case of a negotiation among agents regarding the delivery of information. A promised outcome $o^p \in O^p$ is an assignment of values to the corresponding attribute vector Att^{O^p}, which can be

negotiated by the trustor and trustee. In game theory this kind of non-binding negotiations among agents before the actual interaction takes place is known as "cheap talk" [5].

(b) Effective outcome O^e: The set of attributes Att^{O^e} are not observable until the trust-act has been carried out. Those attributes act as a feedback or judgment for the trustee in respect to his expectations. Att^{O^e} can be thought of as quality aspects of the merchandize, like the delivery time. From a decision theoretic point of view those attributes are the objectives or interests of the trustor and need to be optimized in a multi-criteria optimization problem. From a MAL perspective Att^{O^e} depends on the actions Ac carried out by the opponent.

This way of modeling interaction-trust scenarios allows us to capture almost any context relevant for trust-based decision making.

Our goal is to learn the value function $o^p \rightarrow o^e$ that allows to predict o^e from a given o^p offered by agent a under external conditions c. Moreover, it might be possible to calculate the utility of the trustor for a given o^e. Hence, the ultimate objective is to find the utility function $o^p \rightarrow [0, 1]$. If this function is known the trustor knows what assignment to O^p he should try to achieve (e.g., in a negotiation) to maximize its payoff / reward.

4 Infinite Relational Trust Model

Relational models are an obvious formalization of requirements arising from the relational nature of entities in social, biological, physical and many other fields. The benefits of the relational model for multiagent learning include amongst others:

1. Relational models exhibit flexible and sophisticated modeling capabilities. For typical interaction scenarios in the real world there are more than two types of players and the number of players in each single interaction is flexible or unknown beforehand. In this case propositional models can hardly describe the data and its behavior. Relational models represent correlations both, between the features of an entity and between features of related entities.
2. By transforming the data into a flat representation, also known as propositionalization, the structural information can get lost. Moreover, there is no standard procedure for propositionalization. In [11] manifold propositionalization approaches and their disadvantages are analyzed. In general propositionalization causes high computational costs , since the complexity increases exponentially in parameters, attributes and relations. Another problem of the propositionalization process is the generation of too many irrelevant features. The low quality of propositional features becomes increasingly problematic in more complex interaction models. In contrast, relational models do not need this preprocessing phase at all and do not generate redundant information.

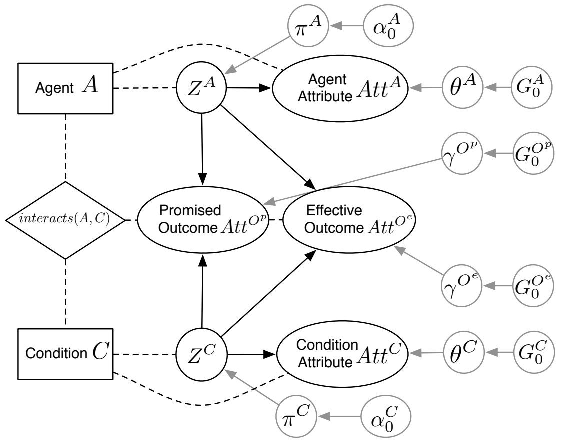


Fig. 1. Infinite Relational Trust Model

3. Relational models and its dynamics can be intuitively visualized by graphical model. This makes complex models more comprehensible and easier to analyze.

Recently [12] and [13] independently introduced infinite relational models (IRM), which express interactions via a potentially infinite number of hidden variables associated with entities instead of difficult structure learning in PRM. Those latent variables play a key role and comprise the inherent structure of the data. As additional features of entities they can improve the accuracy of the learned model.

Considering the properties of nonparametric probabilistic relational models our approach intuitively follows from the interaction scenario that we want to model:

Entity and relationship classes are the two basic building blocks of such a model. In our scenario agents A and states C are both modeled as entities with a corresponding relation $interacts(A, C)$. As a visual representation we make use of the DAPER model (cf. [14]). Figure 1 illustrates the DAPER model for the interaction scenario. Entity classes A and C are depicted as rectangles and the relationship class as a rhombus. Actual evidence *Att* is modeled as attribute classes of entities and relationships (oval). Local distribution classes denoting the parameters and hyperparameters of the probability distributions are shown in small gray circles. The direction of arrows shows the statistical dependency or the sampling process.

The most distinctive feature of our approach are the hidden variables Z (circles). They provide clustering capabilities of entities with a *potentially infinite* number of clusters. Assuming for every entity class one hidden variable our model contains Z^A and Z^C with r^a and r^c clusters, respectively.

4.1 Sampling and Inference

Given the model the essential goal is to infer the conditional distribution

$$P(Z^C, Z^C | Att^A, Att^C, Att^O)$$

of cluster assignments Z^A and Z^C given evidence about relationship attributes Att^{O^p} and Att^{O^e}. This posterior distribution can be formed from the generative models by

$$P(Att_1^{O^e}, ..., Att_k^{O^e}, z_1^A, ..., z_m^A, z_1^C, ..., z_n^C) = \prod_{l=1}^{k} P(Att_l^{O^e} | z_1^A, ..., z_m^A, z_1^C, ..., z_n^C) \prod_{i=1}^{m} P(z_i^A) \prod_{j=1}^{n} P(z_j^C)$$

where we have k actions carried out by m agents and n states. Similar formulas hold for the joint distributions of $P(Att^{O^p}, Z^A, Z^C)$, $P(Att^A, Z^A)$ and $P(Att^C, Z^C)$.

The prior on cluster assignments π^A and π^C is a Dirichlet distribution with hyperparameters α_0^A and α_0^C respectively, where sampling of both Z^A and Z^C can be induced by a Chinese Restaurant Process: $Z|\alpha_0 \sim CRP(\alpha_0)$. By the use of the Chinese Restaurant Process the number of clusters can be determined in an unsupervised fashion. Entities are assigned to (potentially new) clusters corresponding to the size of the existing clusters. Entity attributes Att^A and Att^C are samples from multinomial distributions with parameters $\theta^A \sim G_0^A = Dir(\cdot|\beta^A)$, $\theta^C \sim G_0^C = Dir(\cdot|\beta^C)$ and are generated for each cluster in Z^A and Z^C. The same applies for the relationship attributes Att^{O^p} and Att^{O^e} which can be induced by a multinomial distribution with parameters $\gamma^{O^p} \sim G_0^{O^p}$, and $\gamma^{O^e} \sim G_0^{O^e}$. However, γ needs to be generated for every combination of entity attribute clusters, resulting in $r^A \times r^C$ parameter vectors.

Now inference can be carried out based on Gibbs sampling by estimating $P(Z|Att) \propto P(Att|Z)P(Z)$. For instance the probability of agent i being assigned to cluster k is proportional to $P(z_i^A = k | Z_{j \neq i}^A, Att_i^A, \theta^A, \gamma^{O^p}, \gamma^{O^e}, Z^C) \propto N_k P(Att_i^A | \theta_k^A, \gamma_{k,*}^{O^p}, \gamma_{k,*}^{O^e})$ where N_k is the number of agents already assigned to cluster k and $\gamma_{k,*}$ notes the relation parameters of agent cluster k and all state clusters. Finally, standard statistical parameter estimation techniques can be used for estimating $\gamma_{k^A,k^C}^{O^e}$ from given cluster assignments.

The parameters α_0 and β affect the number of clusters and the certainty of priors and can be tuned. However, we experienced that results were quite robust without extensive tuning. Moreover, our experiments are rather targeted at feasibility than absolute performance, so we fixed $\alpha_0^A, \alpha_0^C = 10$ and $\beta^A, \beta^C = 20$ in all our experiments.

For a detailed description of the algorithm we refer to [12]. We extended the algorithm, as just described, to enable the handling of more than one relationship attribute. Using an arbitrary number of relationships is essential to enable a rich representation of the interaction context and multidimensional trust values.

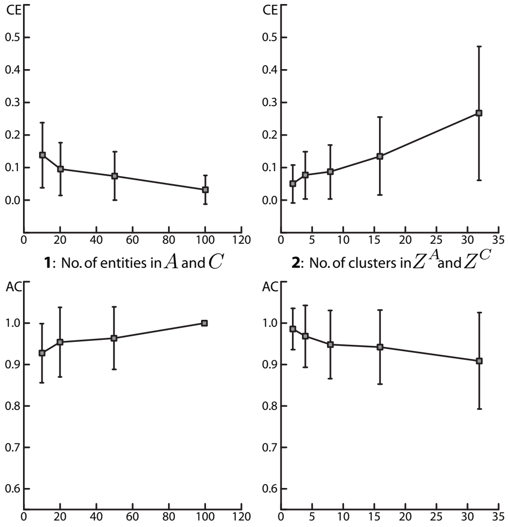


Fig. 2. Results on experiment 1: Synthetic data, setup 1 and 2. Top row graphs show the classification error metric (CE), subjacent graphs show the related accuracy (AC).

4.2 Implications

The ultimate goal of the model is to group entities into clusters Z. A good set of partitions allows to predict the value of attributes Att^{O^p} and Att^{O^e} by their mere cluster assignments. Hereby, our model assumes that each entity belongs to exactly one cluster. It simultaneously discovers clusters and the relationships in-between clusters that are best supported by the data, ignoring irrelevant attributes.

Although the value of attributes is determined entirely by the cluster assignment of associated entities, there is no need for direct dependencies between attributes or extensive structural learning. The cluster assessment of a entity is influenced by all corresponding attributes and cluster assessments of related entities. This way information can propagate through the whole network while the infinite hidden variables Z act as "hubs". As shown in [12] this allows for a collaborative filtering effect. Cross-attribute and cross-entity dependencies can be learned which is not possible with a "flat" propositional approach that assumes independent and identical distributed (i.i.d.) data.

At the same time the number of clusters needs *not* to be fixed in advance. Thus, it can be guaranteed that the representational power is unrestricted.

5 Experiment 1: Synthetic Data

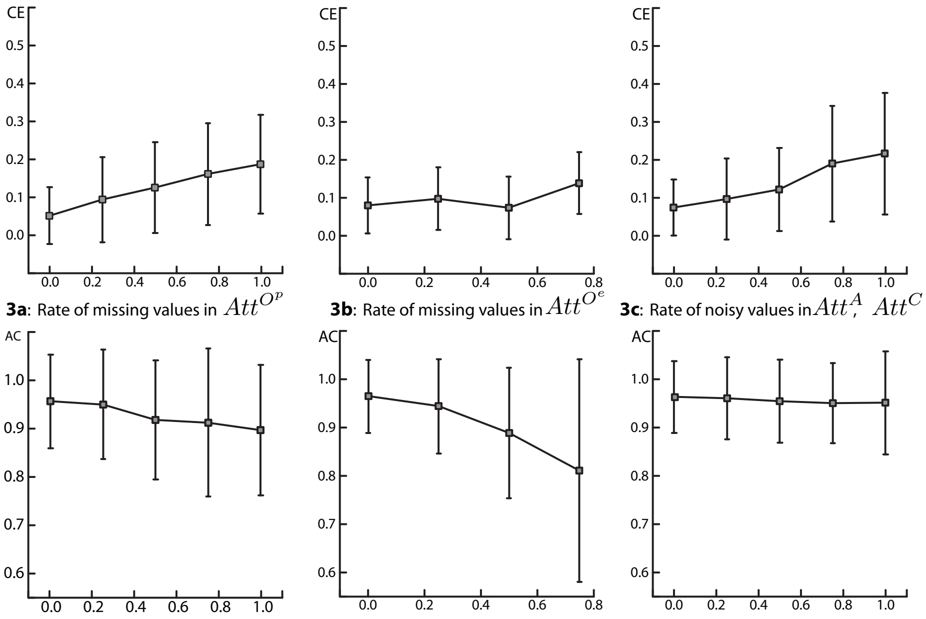
To explore the learning and modeling capabilities of our IRTM we generated synthetic data and evaluated its ability to find clusters in this data. For this purpose we constructed an interaction-trust scenario with the fixed number of 2 entity attributes per entity and 2 relationship attributes, one for O^p and one for O^e. The number of entities $|A|$ and $|C|$ was prespecified but varied in different runs, as well as the underlying clustersize r^a and r^c for Z^a and Z^c. Each entity was randomly assigned to a cluster and its attributes were sampled from a multinomial distribution with 4 possible outcomes and parameter vector θ each. θ in turn, was once randomly generated for each cluster. Accordingly, $r^a \times r^c$ Bernoulli-parameters γ for relationship attribute att^{O^p} and att^{O^e} were constructed.

In Figure 2 and 3 two different error metrics measuring the performance of IRTM averaged over 10 runs are shown. The top row graphs visualize the classification error metric (CE) for clusterings while the bottom row depicts the accuracy (AC) of classifying att^{O^e} correctly. Both are supplemented by a 95% confidence interval. CE reflects the correspondences between the estimated cluster labels and the underlying cluster labels measuring the difference of both (cf. [15]). A value of 0 relates to an exact match, 1 to maximum difference. In this experiment AC is a binary classification task and denotes the ratio of classifying att^{O^e} correctly. Results are averaged over both hidden variables Z^a and Z^c.

5.1 Evaluation

We considered three different experimental setups:

1. We analyzed the performance for different numbers of entities with fixed cluster sizes $r^a = r^c = 4$. The performance shown in Figure 2-1 expectedly suffers for small numbers of entities $|A| = |C| < 20$. Nonetheless, this result suggests that the IRTM is quite robust even with few training samples. This makes it especially interesting for initial trust problems as discussed in the next Section.
2. Correctly recovering different cluster sizes r^a and r^c while the number of entities was fixed to $|A| = |C| = 50$ was the goal of setup 2. In Figure 2-2 we see that the IRTM underestimates the cluster sizes if $r^a = r^c > 16$. This suggests that the number of combinations in such a simple scenario is not enough and entities from different clusters tend to become alike. Still, the AC is almost perfect. Besides that the number of entites per cluster ($|A|/r^a$ and $|C|/r^c$, respectively) gets so small that not all clusters are represented in the training set.
3. Finally, missing and noisy data sets were used in two different ways for training:
   (a) Half of the relationship attribute O^e data was omitted while missing values for O^p was varied. The variance of all measures in figure 3-3a increases with the increase of missing values. Still, the AC is good although cluster correspondences deviate. This clearly shows that dependencies across relationship-attributes have a significant effect on the performance and can be exploited by IRTM. As mentioned before, standard



3a: Rate of missing values in Att^{O^p}

3b: Rate of missing values in Att^{O^e}

3c: Rate of noisy values in Att^A, Att^C

Fig. 3. Results on experiment 1: Synthetic data, setup 3a-c. Top row graphs show the classification error metric (CE), subjacent graphs show the related accuracy (AC).

techniques working with a "flat" vector-based attribute-value representations cannot use such information. In contrast IRTM can propagate information through the network.

(b) First, evidence for O^e was partially omitted. The AC in Figure 3-3b expectedly drops because less training samples of the effective outcome that is to be predicted are available. Still, clustering abilities are hardly affected because other attributes can replace the missing information .

(c) Second, in order to measure the influence of the entity attributes we added noise to Att^A and Att^C. With the used parameter settings IRTM did obviously (see Figure 3-3c) not suffer in predicting AC. However the ability to infer the correct clusters was slightly hindered.

6 Experiment 2: Negotiation Data

Finding an agreement amongst a group of conflicting interests is one of the core issues of distributed Artificial Intelligence. For instance auctions, information markets, preference aggregation and judgement aggregation, game theory and automated negotiations are all research areas that deal with those kind of problems. However most of the approaches neglect the fact that finding the best agreeable solution is not sufficient if the execution of the negotiated outcome can not be enforced by the interaction mechanism. Especially in open systems where agents can enter and leave or change their identity at will, initial trust plays an important role in this regard. The purpose of the IRTM is to make predictions about Att^{O^e} which can be utilized by the agent to adjust its negotiation strategy or trading decisions.

In order to investigate this issue we extended the implementation of a multiagent negotiation framework by an additional trading step. As defined before, let O^p be the promised outcome the agents are negotiating over (e.g., the punctual delivery of information some information agents requested or offered to supply, respectively). This outcome is without loss of generality specified by a set of discrete attributes Att^{O^p}. Now given an assignment of values O^p that two agents have agreed on and promised to fulfill the agents enter an additional trading step where each of them is free to change the assignments of values related to their commitments. Doing so, the agent can decide whether to stick to a bargain or break it at will. One interaction round in this negotiation framework consists of three phases:

1. Negotiation: A strategy that calculates a possible outcome O^p both parties can agree on (e.g., an exchange of goods).
2. Trading: The decision made by every agent whether to stick to a bargain or break it (possibly only partially). The outcomes regarding the agent's obligations are executed according to the agent's decision.
3. Evaluation: The agents can review the effective actions Att^{O^e} of the opponent by observing the received goods and draw conclusions for future interactions

This procedure is repeated over a specified number of rounds with different types of agents.

6.1 Evaluation

Four different agent types were used as opponents in the negotiation game. Every round the negotiation outcome O^p and the effective outcome O^e was recorded. To keep it simple, all agent types follow the same static negotiation strategy but each one acts differently in the trading phase. The agent denoted *Greedy* always maximizes its utility regardless of O^p. *Sneaky*-agent only deviates from O^p if it increases its utility by a large margin, while *Honest*-agent always sticks to O^p. Finally, the agent named *Unstable* deviates only slightly from O^p (by giving away +/-1 amount) if its utility is increased hereby.

As the negotiation strategies were the same for all agent types the negotiation outcome was modeled as attributes of C and not of O^p. Furthermore no specific attributes for A were available except for its identity. Besides the raw negotiation outcome and the state of the own resources, features describing the risk of losing utility and the chance of gaining utility were extracted and added to Att^C. Att^{O^e} was set to be the binary classification task whether the utility would increase less than negotiated or not. This way about 120 interactions were carried out per agent type containing a total of 165 different negotiation outcomes alltogether.

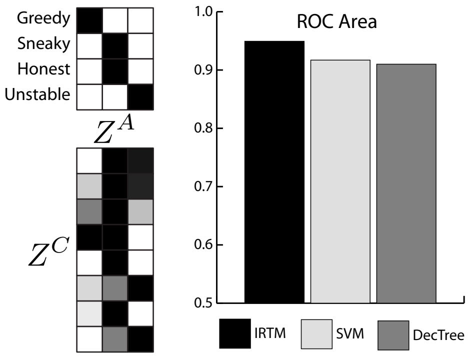


Fig. 4. Results on experiment 2: Negotiation data. Top left shows final clustering of agent types. Bottom left visualizes $P(O^e)$ for each pair of clusters (the darker the more probable). Bar graph shows AUC for classifying $P(O^e)$.

1/3 of the data was randomly withhold and used for testing. Again, all results are averaged over 10 runs.

The predictive performance was measured by calculating the area under the ROC curve (AUC). We compare the results of IRTM to two content based approaches, namely a support vector machine (SVM) using a PolyKernel and a Decision Tree (DecTree, ID3). The SVM and DecTree got an additional input by assigning each agent in A an unique ID number. This way the relational model did not have more information than a "flat" model. We also evaluated the clustering abilities by plotting the most frequent assignment of cluster by the IRTM.

In the top left of Figure 4 one can see that in the end the four agent types (rows) were clustered into three groups in Z^A (columns). Interestingly, the assignment of *Sneaky*- and *Honest*-agent to the same cluster suggests that it is a good strategy to act reliable and provide confidence most of the time in order to convince an opponent of the own trustworthiness. But if it is clear that the gain is really worth it one should betray the opponent's trust.

The rectangles in the lower left corner of Figure 4 visualize $P(O^e|Z^A, Z^C)$. From the 165 different negotiation outcomes and external conditions 8 clusters emerged in Z^C. Each row indicates one condition-cluster Z_i^C, each column an agent-cluster Z_i^A. Thus, each element stands for $P(O^e)$ given the cluster assignments. Brighter rectangles indicate a lower probability for a utility increase as negotiated. As expected the first column (*Greedy*-agent cluster) is on average brighter than the third column (*Unstable*-agent cluster) which in turn is brighter than the middle column (*Sneaky*- and *Honest*-agent).

The overall performance, shown in the bar graph on the right of Figure 4, demonstrates that IRTM has a slightly better performance in classifying $P(O^e)$ than the SVM and the DecTree.

The inherent clustering of the IRTM suggests that it is especially well suited for initial trust situation when unknown but related agents and conditions are

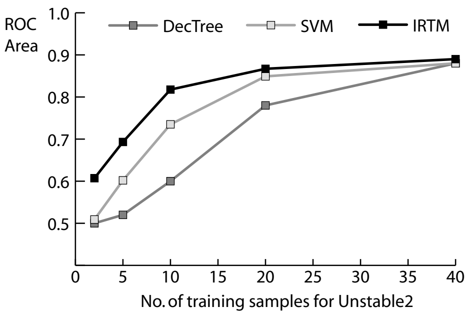


Fig. 5. Results on experiment 2: Negotiation data. Graph shows AUC for different number of samples in an initial trust setup.

observed. Actually, entities can be correctly assigned to a cluster without having seen a single effective outcome related to this entity just by their attributes. To check this assumption we gathered data from interactions with another *Unstable* type agent and evaluated the performance for different numbers of training samples. In the top graph of Figure 5 the AUC is plotted for different numbers of training samples. Especially for a small sample size ≤ 10 the performance of IRTM is clearly better than those of the content based approach.

7 Related Work

As already pointed out, connecting trust to the trusted agent alone without considering contextual and different aspects (dimensions) of trust is not sufficient in many scenarios. Whereas much research on trust concede the importance of context information, most of them do not actually use such information for the calculation of trust degrees in a general and automatic way [16]. To our knowledge using contextual information for initial trust assessment and the transfer of trust between contexts is completely new.

Regarding its dimensionality, most work represent trust as a single discrete or continuous variable associated with one specific agent. Modeling trust in multiple dimensions is only considered by a few elaborate approaches such as [17]. We leave it to the actual scenario how trust needs to be modeled in this respect. In principle, IRTM can handle an arbitrary number of trust variables, each associated with one aspect of the trustor's expectations and represented with any probability distribution needed.

Analogously, we argue that a fine grained modeling of relations between agents and their environment is essential to capture the essence of trust, especially in initial trust situations. There exist a few approaches that can take relationships into account when modeling trust. But in most of this research such relationships are either only considered as reputation or recommendations [18], or as interactions between a group of agents (e.g., [19]). The manifold different kinds of relations that exist between two agents in a specific situational context are not modeled in detail. In addition, most learning techniques are “improvised” for one specific scenario only.

Assessing initial trust values for unknown agents based on pre-specified membership to a certain group has been addressed by [22]. A group-based reputation architecture is proposed here where new agents are assessed according to their pre-specified membership to a certain group of agents. Likewise, the *TRAVOS-C* system proposed by [16] includes rudimentary ideas from hierarchical Bayes modeling by assigning parameter distributions to groups of agents but doesn't come to the point to give a fully automated and intuitive way of how to build clusters.

8 Conclusions and Future Work

In this work, we presented an Infinite Relational Trust Model (IRTM) for interaction-trust and have shown how interactions can be modeled and learned in theory and in two experimental setups. We believe that our model will be especially useful for trust learning in initial trust situations, where the trustor interacts with other agents without having recorded sufficiently enough relevant past experiences in order to judge trustability using traditional methods. E.g., this would typically be the case in short-lived communities of practice, where information agents gather in a kind of ad-hoc manner in order to exchange knowledge, or in open information markets, where mutually more or less unknown information sellers and buyers interact with each other.

IRTM is more powerful and flexible in representing intial trust and fine grained contextual relations, adding a new level of semantics to trust learning. The experimental results suggest that IRTM shows a performance comparable to a “flat” feature-based machine learning approach if trained with independent and identical distributes (i.i.d.) data. We expect to see superior performance of IRTM if no i.i.d. assumption is made and cross-attribute and cross-entity dependencies can be exploited. However, our second experiment shows that in initial trust situations the IRTM can outperform a traditional feature-based approach even if the i.i.d. assumption is made. Besides that, IRTM can handle missing attribute values and enables a clustering analysis which is not possible in existing feature-based trust learning approaches.

Furthermore the experiments deliver preliminary insights into the effect of different strategies on trustworthiness in negotiations. We plan on continuing our work in this direction. Furthermore we intend to address issues like reputation and recommendations which should naturally fit in our relational model.

References

1. Ramchurn, S.D., Hunyh, D., Jennings, N.R.: Trust in multi-agent systems. Knowledge Engineering Review (2004)
2. Willis, J., Todorov, A.: First impressions: Making up your mind after a 100-ms exposure to a face. Psychological Science **17**(7) (2006) 592–598

3. Hu, J., Wellman, M.P.: Multiagent reinforcement learning: Theoretical framework and an algorithm. In Shavlik, J.W., ed.: ICML, Morgan Kaufmann (1998) 242–250
4. Murray, C., Gordon, G.: Multi-robot negotiation: Approximating the set of subgame perfect equilibria in general sum stochastic games. In Schölkopf, B., Platt, J., Hoffman, T., eds.: Advances in Neural Information Processing Systems 19. MIT Press, Cambridge, MA (2007)
5. Littman, M.L.: Markov games as a framework for multi-agent reinforcement learning. In: ICML. (1994) 157–163
6. Wang, X., Sandholm, T.: Reinforcement learning to play an optimal nash equilibrium in team markov games. In S. Becker, S.T., Obermayer, K., eds.: Advances in Neural Information Processing Systems 15. MIT Press, Cambridge, MA (2003) 1571–1578
7. Shoham, Y., Powers, R., Grenager, T.: If multi-agent learning is the answer, what is the question? Technical Report 122247000000001156, UCLA Department of Economics (2006)
8. Brown, G.: Iterative solution of games by fictitious play. In: Activity Analysis of Production and Allocation, New York: John Wiley and Sons (1951)
9. Caruana, R.: Multitask learning. Mach. Learn. **28**(1) (1997) 41–75
10. Krogel, M.A.: On Propositionalization for Knowledge Discovery in Relational Databases. PhD thesis, die Fakultät für Informatik der Otto-von-Guericke-Universität Magdeburg (2005)
11. Xu, Z., Tresp, V., Yu, K., Kriegel, H.P.: Infinite hidden relational models. In: Proceedings of the 22nd International Conference on Uncertainity in Artificial Intelligence (UAI 2006). (2006)
12. Kemp, C., Tenenbaum, J.B., Griffiths, T.L., Yamada, T., Ueda, N.: Learning systems of concepts with an infinite relational model. In: AAAI. (2006)
13. Heckerman, D., Meek, C., Koller, D.: Probabilistic models for relational data. Technical Report MSR-TR-2004-30, Microsoft Research (2004)
14. Meilă, M.: Comparing clusterings: an axiomatic view. In: ICML '05: Proceedings of the 22nd international conference on Machine learning, New York, NY, USA, ACM Press (2005) 577–584
15. Teacy, W.T.L.: Agent-Based Trust and Reputation in the Context of Inaccurate Information Sources. PhD thesis, Electronics and Computer Science, University of Southampton (2006)
16. Maximilien, E.M., Singh, M.P.: Agent-based trust model involving multiple qualities. In: Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS). (2005)
17. Sabater, J., Sierra, C.: REGRET: reputation in gregarious societies. In Müller, J.P., Andre, E., Sen, S., Frasson, C., eds.: Proceedings of the Fifth International Conference on Autonomous Agents, Montreal, Canada, ACM Press (2001) 194–195
18. Ashri, R., Ramchurn, S.D., Sabater, J., Luck, M., Jennings, N.R.: Trust evaluation through relationship analysis. In: AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems, New York, NY, USA, ACM Press (2005) 1005–1011
19. Sun, L., Jiao, L., Wang, Y., Cheng, S., Wang, W.: An adaptive group-based reputation system in peer-to-peer networks. In Deng, X., Ye, Y., eds.: WINE. Volume 3828 of Lecture Notes in Computer Science., Springer (2005) 651–659