Predicting responses of individual reasoners in syllogistic reasoning by using collaborative filtering

Ilir Kola[1] and Marco Ragni[1]

[1] Cognitive Computation Lab, University of Freiburg, 79110 Freiburg, Germany
kola@informatik.uni-freiburg.de
ragni@informatik.uni-freiburg.de

Abstract. A syllogism consists of two premises each containing one of four quantifiers (All, Some, Some not, None) and two out of three objects totaling in 64 reasoning problems. The task of the participants is to draw or evaluate a conclusion, given the premise information. Most, if not all cognitive theories for syllogistic reasoning, focus on explaining and sometimes predicting the aggregated response pattern for participants of a whole psychological experiment. While only few theories focus on the level of an individual reasoner that might have a specific mental representation that explains her response pattern. If different reasoners can be grouped into similar answer patterns then it is possible to identify even cognitive styles that depend on the underlying representation. To test the idea of individual predictions, we start by developing a pair-wise similarity function based on the subjects' answers to the task. For 10% of the subjects, we randomly delete 15% of their answers. By using collaborative filtering techniques, we check whether it is possible to predict the deleted answers of a specific individual solely by using the answers given by similar subjects to those specific questions. Results show that not only the correct answer is predicted in around 70% of the cases, and the answer is in the top two predictions in 89% of the cases, which outperforms other theoretical approaches, but the predictions are as well accurate for cases where participants deviate from the correct answer. This implies that there are cognitive principles responsible for the patterns. If these principles are identified, then there is no need for complex models, because even simple ones can achieve high accuracy. This supports that individual performance in reasoning tasks can be predicted leading to a new level of cognitive modeling.

Keywords: computational reasoning, individual differences, syllogisms, collaborative filtering, machine learning

1 Introduction

Reasoning problems have been studied in such diverse disciplines as psychology, philosophy, cognitive science, as well as in computer science. From an artificial intelligence perspective, modeling human reasoning is crucial if we want to have artificial agents which help us in everyday life. To be successful at this, it is important to understand that each individual can have a different reasoning pattern. Sometimes devia-

tions of the individual participants from the norms of classical logic have led to a qualification of such reasoners as rather irrational (e.g., [27]). Another possibility is that there is a so-called bounded rationality [6]. An indicator could be that these "deviators" are inherently consistent in their answers and even more that their answers can be predicted. Most previous work has focused on overall distribution of answers, trying to predict the most chosen answer by subjects. However, as noted by Pachur and colleagues [18], in presence of individual differences, tests of group mean differences can be highly misleading. For this reason, we focus on individual subjects and try to predict the exact answer they would give.

Collaborative filtering, a method employed in recommender systems [23], can show that a single reasoner does not deviate from similar reasoners, and that consequently her answers can be predicted based on answers of the similar reasoners.

The rest of this paper is structured as follows: first, we give an introduction to theories on reasoning and individual differences, syllogistic reasoning, and recommender systems. Then, we present a model which uses collaborative filtering to predict answers in the syllogistic reasoning task and compare it to other models or theoretical predictions. Lastly, we draw conclusions and suggest further steps for research.

2 Background

2.1 Theories on reasoning and individual differences in reasoning

Scientist have tried to understand human reasoning for a long time. Up to date, there are at least five more prominent theories on how people reason. These theories are based on heuristics [3,4,6,14] mental logic [24,25], pragmatic reasoning schemas [2], mental models [8], and probability theory [15]. Oaksford and Chater [16] offer a general review of these theories.

The need for all these theories is caused by the fact that people differ in how they answer to reasoning tasks. Theories usually aim at explaining general answering patterns, but if we focus on individual answers then these differences are even more vast. These differences can be caused by intellectual abilities, memory capacity, strategies being used, among others [17, 26].

2.2 Syllogistic reasoning

In a syllogistic task, subjects are presented with two premises, and they have to evaluate what follows or whether a third given conclusion necessarily follows. Consider the following example [12]:

Some Artists are Bakers,
All Bakers are Chemists.
Therefore, some Artists are Chemists.

Each premise can have four possible moods, two of which are affirmative (Some, All), and two are negative ones (Some not, No). The premises have two terms each,

but overall only three terms are used. This is because the first two premises always share a common term (in this case bakers), and the third premise asks about the remaining two terms (artists and chemists). Terms can have four figures, based on their configuration:

Figure 1	Figure 2	Figure 3	Figure 4
A-B	B-A	A-B	B-A
B-C	C-B	C-B	B-C

Since each premise can have four moods, and there are four possible figures, there can be 64 distinct pairs of premises. 27 of them have a conclusion which is valid in classical logic, whereas for the remaining 37 there is no valid conclusion. The conclusion (a third statement) allows again four possible moods, and two figures (A-C or C-A), so overall there are 512 syllogisms that can be evaluated.

Studies using syllogisms with different forms of content from abstract to realistic one have shown that errors are not random, but are systematically according to two main factors: figure and mood (see [5]). Syllogistic reasoning has caught the attention of many researchers. Khemlani and Johnson-Laird [12] provide a review of seven theories of syllogistic reasoning. We will describe the ones which perform better in the meta-analysis, and they will be later used as a baseline for the performance of our model.

The first theory, *illicit conversions* [1,22] is based on a misinterpretation of the quantifiers interpreting *All B are A* when given *All A are B* and *Some B are not A* when told *Some A are not B*. Both these conversions are logically invalid, and lead to errors such as inferring All c are a given the premises *All A are B* and *All C are B*. In order to predict the answers of syllogisms, this theory uses classical logic conversions and operators, as well as the two aforementioned invalid conversions.

The *verbal models* theory [20] claim that reasoners built verbal mental models from syllogistic premises and either formulates a conclusion or declares that nothing follows. The model then performs a reencoding of the information based on the information that the converse of the quantifiers Some and No are valid. In another version, the model also reencodes invalid conversions. The authors argue that a crucial part of deduction is the linguistic process of encoding and reencoding the information, rather than looking for counterexamples.

Unlike the previous example, *mental models* (formulated for syllogisms first in [7]) are inspired by the use of counterexamples. The core idea is that individuals understand that a putative conclusion is false if there is a counterexample to it. The theory states that when being faced with a premise, individuals build a mental model of it based on meaning and knowledge. E.g. when given the premise *All Artists are Beekeepers* the following model is built:

Artist Beekeeper
Artist Beekeeper
…

Each row represents the properties of an individual, and the ellipsis denotes individuals which are not artists. This model can be fleshed out to an explicit model which contains information on all potential individuals:

Artist Beekeeper
Artist Beekeeper
Beekeeper

In a nutshell, the theory states that many individuals simply reach a conclusion based on the first implicit model, which can be wrong (in this case it would give the impression that All Beekeepers are Artists). However, there are individuals who built other alternative models in order to find counterexamples, which usually leads to a logically correct answer.

2.3 Collaborative filtering and recommender systems

Recommender systems are software tools used to provide suggestions for items which can be useful to users [23]. One way to implement a recommender system is through collaborative filtering. In a nutshell, collaborative filtering suggests that if Alice likes items 1 and 2, and Bob likes items 1, 2 and 3, then Alice also probably likes item 3. More formally, in collaborative filtering we look for patterns in observed preference behavior, and try to predict new preferences based on those patterns. Users' preferences for the items are stored as a matrix, in which each row represents a user and each column represents an item. Then, for each user we build a similarity function to see who are the users which have similar preferences. This means, for each user we have a neighborhood of other users similar to them. Then, when a certain item has not been rated by our user, we rely on this neighborhood to see how would our user rate that item. If the rate would be high enough, we can recommend that item to the user.

$$\mathbf{R} = \underbrace{\left.\begin{bmatrix} ? & ? & 1 & \dots & 4 \\ 3 & ? & ? & \dots & ? \\ ? & 5 & ? & \dots & 5 \end{bmatrix}\right\} U \text{ users}}_{M \text{ items}}$$

Fig. 1. Users' ratings represented as a matrix

The main challenge in this case would be to select the appropriate similarity function, and to determine the adequate size of the neighborhood.

3 Predicting performance in syllogistic reasoning by using collaborative filtering

3.1 Motivation

As aforementioned, people make mistakes when solving reasoning tasks such as syllogisms. When it comes to preference behavior, we have seen that collaborative filtering can achieve very good results in predicting which items to recommend to users. This shows that people are consistent in their preferences. Could it be the case that people are also consistent in the way they perform in reasoning tasks, and can we predict their answers (including errors) in the aforementioned reasoning domains? We will explore this by using collaborative filtering to predict participants' behavior in reasoning tasks.

3.2 The experimental setting

For this model, we will use an unpublished data set from an online experiment conducted at the Cognitive Computation Lab (University of Freiburg). It includes data from 140 subjects which completed all 64 syllogistic tasks. Each subject was presented with two premises, and had to choose between nine answer options (the eight mood/figure combinations, plus the ninth option being No Valid Conclusion).

3.3 The model

In our setting, the users are the 140 subjects of the study, and the items are the 64 tasks. We define the similarity function as follows:

$$sim = \frac{n_{sameAnswers}}{N} \quad (1)$$

where N represents the amount of questions which were answered by both subjects. As we can see, similarity is a function between 0 and 1.

We start by randomly selecting 14 subjects for which there exists at least one other subject with a similarity of 0.6 or higher, and then randomly deleting 10 of their answers. These will be the answers which have to be predicted.

The model computes the pair-wise similarities between subjects, and then whenever for the current subject there is a missing answer, it identifies all subjects in its neighborhood (i.e., subjects with a similarity higher than 0.35) which have answered that task, and performs a "weighted voting" as following:

```
for answer in possible_answers:
   for user in users:
      value[answer]=value[answer]+sim[user]*given[user]
```

where `sim[user]` represents the similarity of our subject with the user which we are currently computing, and `given[user]` is a binary attribute showing whether

the user gave this answer to the task or not. We perform this weighting inspired by the intuition that answers given by more similar subjects should matter more. Then, the answer with the highest value is the predicted one.

3.4 Results

The model is very simple, and it does not include any learning, its performance is, however, fairly accurate. It is important to notice that the model predicts one out of nine possible options, so a model which is simply guessing would be on average correct in about 11% of times. Our model compares the predicted answers to the true ones, and reports the percentage of correctly predicted answers.

In order to interpret the result better it would be useful to compare the performance of our model to other models or theoretical predictions. As we already stated, most theories do not focus on individual answer predictions, but on most chosen answers. For example, a theory can state that for the premises *All A are B*, *Some B are C* then people draw the answers *Some A are C*, *Some C are A* or *All A are C*. We try to see what these theories would predict for our individual missing answers, and we use the relaxation that if the missing answer is one of the predicted answers from the theory, then it is counted as correct. We notice that this is quite a big relaxation, since there are theories which predict three to four answers for the same pair of premises, which means they would of course achieve a better accuracy than our model which always predicts just one answer. We calculate the accuracy of the predictions of theories based on illicit conversions, verbal models, and mental models, as well as the predictions of mReasoner, an implementation of the mental models theory.

One thing to keep in mind is that for some syllogisms there is more than one valid answer, however subjects could select in our experiment only one answer. This can cause a difficulty for our comparison as we need to deal with cognitive theories that often predict up to four or five answers per syllogism. For this reason, we construct two other versions of our model. Instead of predicting only one answer, we checked what would be the accuracy of the prediction if we predict the top two and top three most voted answers. We repeat the procedure 100, 300 and 500 times (to check if results converge); the results are reported in Table 1:

	Exact	Top 2	Top 3	IC	VM	MM	mReasoner
100 runs	0.68	0.89	0.95	0.61	0.77	0.95	0.87
300 runs	0.69	0.88	0.95	0.62	0.77	0.95	0.87
500 runs	0.69	0.89	0.95	0.62	0.77	0.95	0.87

Table 1. The accuracy of the cognitive theories in predicting missing answers. The reported results are average accuracies over 100, 300, and 500 runs. (Exact, Top2 and Top3 refer to our model producing 1, 2 and 3 answers; IC refers to Illicit Conversions; VM refers to Verbal Models; MM refers to Mental Models)

3.5 Discussion

The results show that our model which predicts the exact answer does not only perform reliably better than chance, but even manages to outperform the theoretical predictions based on illicit conversions, which for almost half of the syllogisms predicts more than one answer. Furthermore, we notice that our model with the two most voted predictions outperforms the predictions of the verbal models as well as of mReasoner, which is right now one of the state-of-the-art predictors for syllogistic reasoning. Another thing which is important to notice is that our model reaches the same accuracy even if we delete 32 (out of the 64) answers for up to 50% of the participants, showing robust performance.

We notice that the top performance is achieved by the predictions made by the mental models theory. However, it is important to notice that for almost half of the syllogisms this theory predicts four or even five answers, which means it has an advantage for this type of metric. Still, our model which predicts the top 3 answers (still less than the mental models predictions) achieves the same performance.

mReasoner is an implementation which is based on mental models, but it has some parameters which limit the number of predicted answers for each syllogism (it predicts one answer for 7 syllogisms, and more than two for 16 syllogisms). In this comparison, we used the default setting for mReasoner, and we see that our model which predicts the top two answers has a better performance.

Khemlani and Johnson-Laird [13] propose a model where mReasoner learns parameters for individual subjects in a small dataset consisting of 20 participants, and then simulates the answers of each subject and compares them to the true answers. They report a mean correlation to the data of 0.7, which means on average in 70% of the cases mReasoner made the right prediction. This result is comparable to our basic model, but built on general cognitive principles. Both approaches differ in their methodology: Our approach requires participants data to classify and predict other reasoners and does not have cognitive principles, while on the other hand mReasoners is built on cognitive principles but trains the system parameters on the whole dataset, so it is not actually predicting. A combination of both methods to reach a "prediction" based on cognitive principles is important.

4 Conclusions and future steps

These results show that collaborative filtering can help in predicting individual performance for reasoning tasks, but also that there are new challenges (especially by the performance boost when considering the top two predictions). First of all, it will be interesting to test the same model with data from other reasoning domains, e.g., the Wason selection task [28]. This would allow us to test for consistency across different reasoning domains. Secondly, as we mentioned the model is simple, it would be interesting to build a more adaptive model which learns from the subjects' answers and can identify cognitive principles. This could be achieved by analyzing potential rea-

sons for differences in performance, combined to using more advanced techniques from machine learning to build the recommender system.

One alternative would be to formalize the tasks by using ternary logic, and then learn how different subjects map logical operators to truth tables. Ternary logic has shown to provide high flexibility in modeling Wason's selection task [21]. Another alternative would be to include theories' predictions to the task, and check whether a subject is consistent with the predictions of a certain theory (i.e. we would find similarities with theoretical predictions rather than with other subjects). This would also help us for cases where there are not enough subjects to build informative similarity functions among them.

We tried to use machine learning techniques to cluster the data in order to identify potential reasoning profiles, however the dataset seems to be too diverse. A method called fcclusterdata, a hierarchal clustering technique from the sckit-learn package [19] in Python, identifies more than 40 clusters (for the 140 participants), whereas by using the k-medoids technique, in which we can specify the number of clusters, for up to 6 clusters the similarity of subjects in the cluster remains low and we do not achieve better performance. Studies [17, 26] have identified reasons which might lead to individual differences such as level of intellect, memory capacity etc. Our intuition is that although these reasons are similar for different individuals, the way they are presented in people makes it difficult to create clusters. For example, an individual can have high intellectual capacity but bad memory, another one medium intellectual capacity and very good memory, and so on. This is why we think that an approach which focuses on finding similar reasoners for each individual can be more effective.

Reasoners are relatively consistent in their performance in syllogistic reasoning, since some tend to give similar answers and often predictable mistakes. This means it is possible to build reasoning models which can identify a person's reasoning pattern, and exploit it to better understand the overall reasoning process. This is exactly what our simple model does, and in its relaxed version it manages to be as good as state of the art complex reasoning models.

References

1. Chapman, L. J., & Chapman, J. P. Atmosphere effect re-examined. Journal of experimental psychology, 58(3), 220. (1959)
2. Cheng, P. W., & Holyoak, K. J. Pragmatic reasoning schemas. Cognitive psychology, 17(4), 391-416. (1985)
3. Evans, J. St. B. T. Heuristic and analytic processes in reasoning. British Journal of Psychology, 75, 451-468. (1984)
4. Evans, J. St. B. T. Bias in human reasoning: Causes and consequences. Hillsdale, NJ: Erlbaum. (1989)
5. Evans, J. S. B., Newstead, S. E., & Byrne, R. M. Human reasoning: The psychology of deduction. Psychology Press. (1993)
6. Gigerenzer, G., & Hug, K. Domain-specific reasoning: Social contracts, cheating, and perspective change. Cognition, 43(2), 127-171. (1992)
7. Johnson-Laird, P. N. Models of deduction. Reasoning: Representation and process in children and adults, 7-54. (1975)

8. Johnson-Laird, P. N. Mental models: Towards a cognitive science of language, inference, and consciousness (No. 6). Harvard University Press. (1983)
9. Johnson-Laird, P. N., & Steedman, M. The psychology of syllogisms. Cognitive psychology 10.1: 64-99. (1978)
10. Johnson-Laird, P. N., & Wason, P. C. A theoretical analysis of insight into a reasoning task. Cognitive Psychology, 1(2), 134-148. (1970)
11. Kaufman, L., & Rousseeuw, P. J. Finding Groups in Data: An Introduction to Cluster Analysis, Wiley New York Google Scholar. (1990)
12. Khemlani, S., & Johnson-Laird, P. N. Theories of the syllogism: A meta-analysis. Psychological Bulletin, Vol 138(3), May 2012, 427-457. (2012)
13. Khemlani, S., & Johnson-Laird, P. N. How people differ in syllogistic reasoning. In Proceedings of the 36th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society. (2016)
14. Newell, A., & Simon, H. A. Human problem solving (Vol. 104, No. 9). Englewood Cliffs, NJ: Prentice-Hall. (1972)
15. Oaksford, M., & Chater, N. A rational analysis of the selection task as optimal data selection. Psychological Review, 101, 608-631. (1994)
16. Oaksford, M., & Chater, N. Theories of reasoning and the computational explanation of everyday inference. Thinking & Reasoning, 1(2), 121-152. (1995)
17. Oberauer, K., Süß, H. M., Wilhelm, O., & Sander, N. Individual differences in working memory capacity and reasoning ability. Variation in working memory, 49-75. (2007)
18. Pachur, T., Bröder, A., & Marewski, J. N. The recognition heuristic in memory-based inference: is recognition a non-compensatory cue? Journal of Behavioral Decision Making, 21(2), 183-210. (2008)
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830. (2011)
20. Polk, T. A., & Newell, A. Deduction as verbal reasoning. Psychological Review, 102(3), 533. (1995)
21. Ragni, M., Dietz, E. A., Kola, I., & Hölldobler, S. Two-Valued Logic is Not Sufficient to Model Human Reasoning, but Three-Valued Logic is: A Formal Analysis. Bridging 2016 – Bridging the Gap between Human and Automated Reasoning, 1651:61–73. (2016)
22. Revlis, R. Two models of syllogistic reasoning: Feature selection and conversion. Journal of Verbal Learning and Verbal Behavior, 14(2), 180-195. (1975)
23. Resnick, P., & Varian, H. R. Recommender systems. Communications of the ACM, 40(3), 56-58. (1997)
24. Rips, L. J. Cognitive processes in propositional reasoning. Psychological review, 90(1), 38. (1983)
25. Rips, L. J. The psychology of proof: Deductive reasoning in human thinking. MIT Press. (1994)
26. Stanovich, K. E., & West, R. F. Individual differences in rational thought. Journal of experimental psychology: general, 127(2), 161. (1998)
27. Tversky, A., & Kahneman, D. Availability: A heuristic for judging frequency and probability. Cognitive psychology, 5(2), 207-232. (1973)
28. Wason, P. C. Reasoning. New Horizons in Psychology. pp. 135-151. (1966)