Towards Enabling Crowdsourced Collaborative Data Analysis

Feldman, Michael ; Anastasiu, Cristian ; Bernstein, Abraham

Towards Enabling Crowdsourced Collaborative Data Analysis

MICHAEL FELDMAN, University of Zurich
CRISTIAN ANASTASIU, University of Zurich
ABRAHAM BERNSTEIN, University of Zurich

1. INTRODUCTION

The availability of data is growing faster than the availability of experts with the relevant skillset for interpreting it. Finding competent experts for data analysis tasks is becoming increasingly challenging due to the variety of required skills. Business and academic settings assume analysts to be proficient not only in the domain of their interest, but also in core analysis disciplines such as statistics, computing, software engineering, and algorithms. Data analysis routines in these domains span over multiple disciplines and most people approach them with their own set of biases as well as limited knowledge potentially leading to errors (MacCoun 1998).

Motivated by this challenge, this study explores the idea of collaborative data analysis, where it is assumed that every member of an analysis team possesses a tiny fragment of the required knowledge and, taken together, they can use their collective intelligence for successful data analytics (Bernstein et al. 2012). Specifically, *we propose and evaluate an approach to process complex data analysis inquiries with the aid of lay statisticians and enthusiasts possessing only limited knowledge about data analytics*. This paper proposes a collaborative data analysis framework allowing structured data analysis tasks to be distributed as a collaborative process to a group of people with a diverse set of skills and knowledge. The proposed approach is examined through two hypotheses: (a) data analysis projects can be decomposed into small enough tasks such that non-experts can successfully perform on them and (b) teams with a mixed level of expertise perform as well as standard expert based projects. Our evaluations showed that data analysis tasks, with a focus on the *pre-processing activities*, might be successfully distributed and accomplished by the non-expert workers. Moreover, the outputs of the crowdsourced data analysis are equivalent in quality and competitive in cost in comparison with the expert-based work.

2. EXPERIMENTAL DESIGN

Platform choice: We have analyzed different frameworks and tools for addressing the collaborative data analysis in crowdsourcing settings. We evaluated the platforms for ease of use, scalability, extendibility, open source license, straightforward results communication, as well as coordination and collaboration features. To support collaboration of non-experts, different alternatives such as Git, Veracity, and DataHub (Bhardwaj et al. 2015, Kandogan et al. 2015) were reviewed. Given that none of these systems fulfilled all our requirements, we decided to implement our prototype with aid of Jupyter Notebook (IPython) – a client/server application that allows editing and running notebooks (i.e., descriptions of computations) in a web browser either locally without internet access or installed on a remote server. Using Jupyter, researchers can capture data-driven workflows that combine code, equations, text, and visualizations and share them with others. Our decision to adopt this tool was guided by the following considerations: (a) Jupyter is browser-based notebook with support for code, text, mathematical expressions, and inline plots, (b) although initially designed for Python, it is language agnostic and can process languages such as R, Ruby and others, and (c) it supports interactive data visualization toolkits, frequently required in data analysis.

Collaboration extensions: We based our system on a Jupyter extension that allows using Google Drive for file management and supports contributors with access over a web interface, and other features such as sharing notebooks and adding different sets of permissions. In addition, we had to develop the following features to support collaborative data analysis: (1) workflow guidance for creating data analysis/mining projects and distributing tasks to different users, (2) an environment to manage and modify projects, (3) the ability for users/collaborators to reflect and comment on their colleagues' notebooks akin to Microsoft Word's comment/reviewing feature, (4) functionality to merge all notebooks of a project into a master notebook, which can run all the different distributed steps in one run, and (5) the ability to have an iterative collaboration process.

The project creation workflow was designed to allow a data analysis expert, who will act as the project manager, to define project and distribute actions (or assignments) to workers in a top-down approach. Splitting tasks into small actions allows the project manager to group and distribute them to non-expert workers. The assigned workers then perform their assigned actions (maybe commenting on some other elements of a notebook). The manager then integrates the elements, potentially deciding to iterate on some of them.

In our framework, an *action* is the smallest unit into which a task can be split and assigned. It is described by its name, input, and output. The project manager can assign the actions to different contributors according to the considerations such as required expertise, worker availability, or interdependency of actions. An example of an action is *loadDFFromCSV*, which receives as input the path of the CSV file and returns a data frame. The project owner can search or filter for actions from a default taxonomy and group them into assignments. Specifically, we picked the taxonomy compiled of the "Catalogue of Methods in Data Pre-Processing" created by AixCAPE e.V.[1] and methods proposed by the Salvador et al. (2015).

The prototype was then used in three experiments to evaluate our hypotheses.

3. EVALUATION

To test the hypotheses we carried out experiments with crowd workers and compare the results with those of experts.

Experimental tasks: We used the crowdsourcing platform Upwork to recruit the target user group for our system — non-expert data scientists. To find the projects to be crowdsourced we used the data science platform Kaggle — a well-known data science platform. Given the high quality of many winning Kaggle analyses, we assume that these projects were solved by data science experts, and use their results for comparison with the results of the crowd workers. We compared the quality of the output crowd's performance when employing our prototype with winning/published Kaggle results by correlating the results, testing for equivalence with confidence intervals (Mascha 2010) and by constructing a Bland–Altman plot (Bland and Altman 1986). Keeping in mind the phenomenon where different data analysts are using different methods on the same dataset to answer the same question end up with a wide variety of conclusions (Silberzahn and Uhlmann 2015), we tried to choose projects with most objective outcomes that underwent through public review on Kaggle.

The Kaggle projects we chose were "US Census,"[2] "Hillary Clinton's Emails,"[3] and "Reddit Sentiment Analysis."[4] The goal of the first project was to create a chart showing the earnings of the population by occupation and gender. The main focus is on finding the right occupation categories and sub-setting the data accordingly. The goal of the second project is to create a heat map based on the frequency the countries are mentioned in the emails sent by Hillary Clinton. In the last project

[1] http://dataprocessing.aixcape.org/index.php/Single_steps

[2] https://www.kaggle.com/wikunia/d/census/2013-american-community-survey/earnings-by-occupation-sex

[3] https://www.kaggle.com/ampaho/hillary-clinton-emails/foreign-policy-map-through-hrc-s-emails/code

[4] https://www.kaggle.com/lplewa/d/reddit/reddit-comments-may-2015/communication-styles-vs-ranks

crowdworkers were asked to create a chart showing which Reddit comments receive the highest scores, based on the sentiment of the comment.

Results: Our first hypothesis proposes that *(H1): the pre-processing part of a data analysis project can be decomposed into small enough tasks such that non-experts —workers with limited coding skills and no or basic data analysis skills — can successfully perform on them*. We found that a non-expert data analyst (recruited locally) could split all three experimental projects into actions and assign them to workers. We also found that a similar decomposition could be found in the solutions presented on Kaggle. The Upwork workers were able to successfully complete their assignments and rated the complexity of their assignment an average of 2.1 out of 5, which is lower than the average project complexity at 2.3 out of 5. Even though these results should be taken with the grain of salt, given the small sample of workers ($n= 9$), this indicates that we were able to divide a slightly complex project into several less complex assignments. Obviously, more experimentation is needed to support this hypothesis.

The second hypothesis states that *(H2): the proposed team with a mixed level of expertise performs as well as standard expert-based projects*. In order to test this hypothesis, we analyzed two elements of this project. First, we evaluated our collaboration tool and workflow via a survey. The tool was used by a group of hired crowdworkers that were asked to conduct data analysis projects, in order to see whether the equivalent quality of results can be achieved using our tool versus traditional development tools usually used by experts. We examine whether the tasks can be decomposed with relative ease, whether the projects can run faultlessly and, eventually, we collect feedback on the platform from the participants. Additionally, we also evaluate the quality of the crowdsourced projects. The preliminary analysis of the survey indicates that most crowdworkers liked the tooling (rating is at 3.9 out of 5), were able to communicate using the tool, and did not rely on any additional software. Second, the results produced by the crowd workers and Kaggle are highly correlated (0.8 and 0.72 for the first and second experiments respectively). A t-test conducted on the third experiment does not support the hypothesis of true difference in results. Moreover, the results reside within Bland-Altman plot and equivalence test shows no significant difference between the results of expert and non-expert teams (i.e., less than ¼ of the pooled standard deviation of the compared results). Hence, the evaluation shows the potential of our proposed approach to enable non-expert crowd workers to collaboratively produce expert-like results when guided by a project manager and encourages future research.

CONCLUSION

The goal of this paper is to contribute to the field of data analysis by discussing and implementing a framework that allows for collaborative data analysis in crowdsourcing environments. We created an online platform that permits the efficient splitting of data pre-processing into multiple tasks. These tasks can then be assigned to crowdworkers/freelancers with little to no expertise on the subject matter of the overall project, but who can solve smaller, simpler assignments (e.g. using basic coding skills). We tested our tool with three projects accomplished by teams of non-experts. Our experiments showed that the guided non-experts generated results comparable to the ones produced by experts. Furthermore, in our experiment the total cost was about 120 USD per project (the projects were split between three crowdworkers), where every worker has been paid 40 USD to accomplish her part and each project required on average about 12 hours of work. This makes the projects economically competitive with expert-based projects, especially in the light of the soaring data scientists' salaries. To conclude, we present a proof of concept prototype for collaborative data pre-processing and analysis. We also illustrate that through supporting the collaboration of non-expert users they can be successfully included in more complex data analysis projects, producing outputs comparable in quality and at a lower cost than expert data scientists.

REFERENCES

Abraham Bernstein, Mark Klein, and Thomas W. Malone. 2012. Programming the global brain. *Communications of the ACM* 55.5 (2012), 41-43.

Anant Bhardwaj, Amol Deshpande, Aaron J. Elmore, David Karger, Sam Madden, Aditya Parameswaran, Harihar Subramanyam, Eugene Wu, and Rebecca Zhang. 2015. Collaborative data analytics with DataHub. *Proceedings of the VLDB Endowment* 8, 12 (2015), 1916-1919.

J. Martin Bland, and Douglas G Altman. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet* 327.8476 (1986), 307-310.

Eser Kandogan, Mary Roth, Peter Schwarz, Christina Christodoulakis, and Renée J. Miller. 2015. LabBook: Metadata-driven social collaborative data analysis. In *Proceedings of the Big Data IEEE International Conference (Big Data)*. IEEE, Santa Clara, CA, 431-440.

Robert J. MacCoun. 1998. Biases in the interpretation and use of research results. *Annual review of psychology* 49. 1 (1998), 259-287.

Edward J. Mascha. 2010. Equivalence and noninferiority testing in anesthesiology research. *The Journal of the American Society of Anesthesiologists* 113,4 (2010), 779-781.

García Salvador, Julián Luengo, and Francisco Herrera. 2015. *Data preprocessing in data mining*. Switzerland: Springer, 2015.

Raphael Silberzahn and Eric L. Uhlmann. 2015. Crowdsourced research: Many hands make tight work. *Nature* 526, 7572 (2015), 189.