INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

Crowdsourcing: A Result Analysis

Jayshri N. Ganthade
Professor, Amrutvahini College of Engineering, Pune University), Sangamner, India
Dr. S. R. Gupta
Professor, Ram Meghe Institure & Research Center, Badnera, Amravati, India

Abstract:
Models are generating from large data sets— and determining which subsets of data to mine— are becoming increasingly automated. What data to collect in the first place requires human intuition or experience, usually choose and supplied by a domain expert. In this paper a new approach is described to machine science which demonstrates for the first time that non-domain experts (crowd) can collectively formulate features, and provide values for those features and do the result analysis, which are based upon the result which will be output of this research. This was accomplished by building a web platform in which human groups should be interact to respond to questions likely from which behavioral outcome will be predict and pose new questions to their database if interested.
Here two web-based experiments have described, in the first site led to models that can predict users' monthly electric energy consumption; the other led to models that can predict users' body mass index. The values which are entered by user for Energy consumption and body mass index are used to analysis how the energy consumption can be reduced and how the body mass index should be maintained.

Keywords*: Crowdsourcing, machine science, Body Mass Index, energy consumption*

1. Introduction

Multiple regression or neural networks which provide mature methods for computing model parameters are the statistical tools when the set of predictive covariates and the model structure are pre-specified.
However, the task of choosing predictive variables for study is largely a qualitative task that requires substantial domain expertise. For example, a designer of survey must have the experts to choose questions that will identify predictive covariates. An engineer must improve substantial familiarity with a design in order to determine which variables can be systematically adjusted in order to optimize performance.
The need for the involvement of domain experts can become a bottleneck. However, if the wisdom of crowds could be harnessed to produce insight into difficult problems, one could be find out exponential rises in the discovery of the causal factors of behavioral outcomes, mirroring the exponential growth on other online collaborative communities. Thus, the goal of this research was to test an alternative approach to modeling in which the wisdom of crowds is harnessed to both propose potentially predictive variables to study by asking questions, and respond to those questions, in order to develop a predictive model.

1.1. Machine science

The explosion of knowledge is changing the landscape of science. Now a day, Computers play an important role in helping scientists to store, manipulate, and analyze data. New capabilities are extending the reach of computers from analysis to hypothesis. Drawing an approach from artificial intelligence, computer programs are able to integrate published knowledge with experimental data and enable new hypotheses to emerge with little human intervention.
Recent research demonstrated that how scientists can be used computers to become better informed and more graceful explorers. The pool of concepts and relations which are used for generating automated can be expanded by the new computational tools that can be hypotheses by (i) drawing more from the vast collection of written texts of published science, and (ii) synthesize the new higher order and lower order concepts and relations from the existing knowledge[1]

1.2. Crowdsourcing

On the Internet, the rapid growth in user-generated content is an example of how bottom-up interactions, under some circumstances, can effectively solve problems that previously required explicit management by teams of experts [2]. "crowdsourcing" is nothing but it is the harnessing the experience and effort of large numbers of individuals and has been used effectively in a number of research and commercial applications [3].In this crowdsourcing tool a "Human Intelligence Task" such as characterizing data [4], transcribing spoken language [5], or creating data visualizations [6] are described by human. it involves large groups of humans from many locations it is possible to complete tasks that are difficult to accomplish with computers alone, and expensive to accomplish through traditional expert-driven processes [7].

The problem solving through crowdsourcing can produce novel, creative solutions that are substantially different from those produced by experts. An iterative, crowdsourced poem translation task produced translations that were both surprising and preferable to expert translations [8]. We conjecture that crowdsourcing the selection of predictive variables can reveal creative, unexpected predictors of behavioral outcomes. For problems in which behavioral change is desirable (such as is the case with obesity or energy efficiency), identifying new, unexpected predictors of the outcome may be useful in identifying relatively easy ways for individuals to change their outcomes.

2. Methodology

The system described here modeling a human behavior such that: (1) the administrator (investigator) defines some human behavior-based outcome that is to be modeled; (2) data is collected from crowd(human intelligence task); (3) models are continually generated automatically; and (4) the crowd is motivated to propose new independent questions which are used for further applications[9].

Fig. 1 illustrates how the investigator (administrator), participant group (crowd) and modeling engine work together to produce predictive models of the outcome. The administrator begins by constructing a web site and defining the human behavior outcome to be modeled. In this paper a energy and BMI outcome were investigated: electric energy consumption of an individual owner, and their body mass index. The investigator then initializes the site by seeding it with a small set (three or four) of questions known to correlate with the outcome of interest. For example, based on the suspected link between electricity consumption and electricity bill, we seeded the Energy website with the question "do you have electric clothes dryer?"

2.1. Algorithm

The algorithm requires two inputs such as Q (the set of initial seed questions) , Φ (the set of post seed questions) the questions which are posed by the users and one output set ,P.

The questions are initialized by the investigator. The user who has visited the site, first select the survey name. Answer the questions which are displayed in the survey .The user may post some questions with the answers of these questions. The investigator adds these questions (qi) for the next survey (Si), which is answered by the user. If the questions are valid which is posed by the users are also added to Si (for survey questions). Finally the model will be generated (P).

Input: $Q=\{q_1,q_2,\ldots\ldots,q_n\}$ Initial seed question
$\Phi=\{\Phi_1,\Phi_2\ldots\ldots\ldots,\Phi_n\}$. post seed question

Output: $P=(Q \cap \Phi)$ when 't' satisfy, generate model

Process:

Step 1: Initialize initial seed questions (by investigator)
$Q=\{q_1, q_2,\ldots\ldots,q_n\}$

Step 2: Enter survey name (by user)

Step 3: Enter the value for corresponding survey (BMI or Energy consumption)

Step 4: Start survey process (by the user).

Step 5: add q_i to S_i

Where S_i be the survey to which q_i has to be added (answered by the user).

Step 6: wait till Φ_t = true.

Step 7: while (Φ)

```
{
If (Φ_i satisfy i. e. true)
{
P=Q ∪ Φ;
}
}
```

Step 8: repeat Step 5

Step 9: generate model.

Step 10: end.

As shown in the figure 1, when there is user, he has to select survey as BMI or energy consumption. In both the cases it is restricted to enter BMI or energy consumption respectively. After providing these values, visit to respective survey. Answer all the questions if possible. There is no restriction to answer all the questions. Submit the survey. If the user wants he can submit his own questions, which will be added to survey after the decision of investigator. Logout from survey.
If there is Administrator, there are various options available. Survey Manager will be adding survey. Question manger will be viewing the questions, adding own questions. Investigator having the rights to add or delete the questions which are directly added to the database. Model is used to generate model which is nothing but the results. Before generating the model, the administrator has to select which model is to generate .Finally the administrator will be Logout.

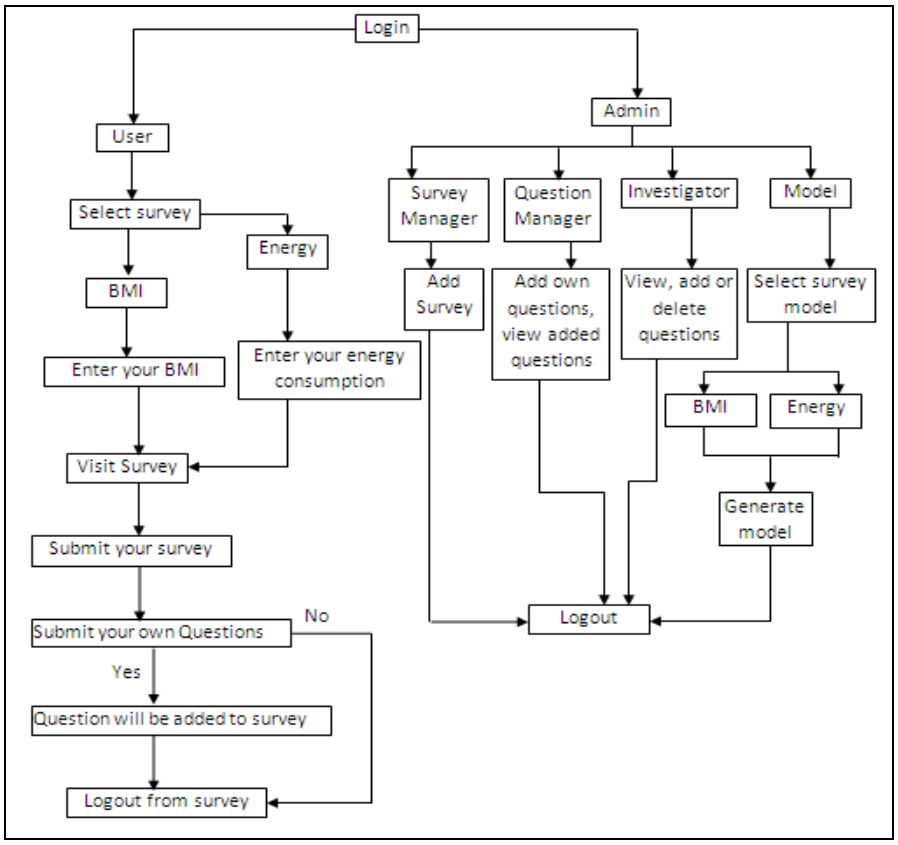
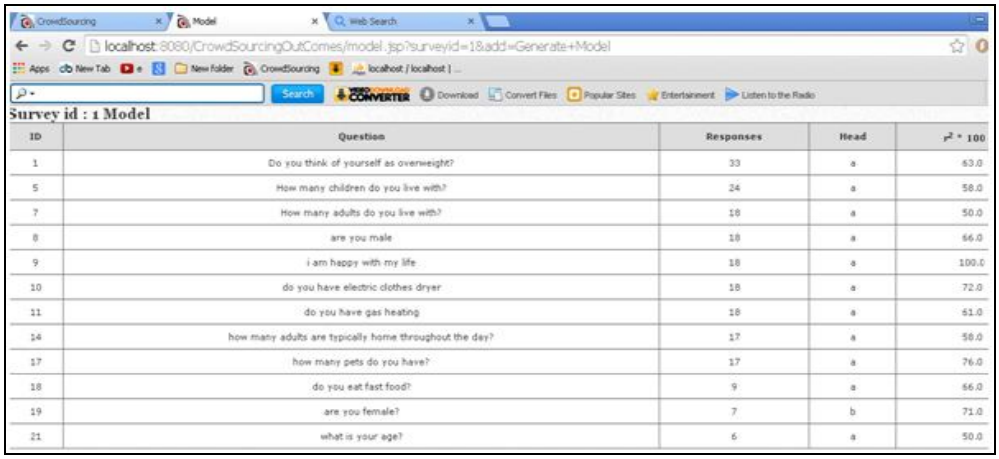


Figure 1: Data Flow diagram

3. Result analysis


Survey id : 1 Model

ID	Question	Responses	Head	r² * 100
1	Do you think of yourself as overweight?	33	a	63.0
5	How many children do you live with?	24	a	58.0
7	How many adults do you live with?	18	a	50.0
8	are you male	18	a	66.0
9	i am happy with my life	18	a	100.0
10	do you have electric clothes dryer	18	a	72.0
11	do you have gas heating	18	a	61.0
14	how many adults are typically home throughout the day?	17	a	58.0
17	how many pets do you have?	17	a	76.0
18	do you eat fast food?	9	a	66.0
19	are you female?	7	b	71.0
21	what is your age?	6	a	50.0

Figure 2: Model generated for Body Mass Index (BMI)

Above Figure 2 shows the result of Body Mass Index. We are having the column of questions, number of responses, head and response rate .From above table , we should know that what should be question, what are responses for particular questions, what should be the expected answer for particular question , what should be the response rate.
Similarly, Figure 3 shows the result of Energy consumption. Here also we are having the different columns such as question, responses, head and response rate $r^{2.}$ From above these two figures (Figure 3 and Figure 4), we should find out the relations between the various field of the tables.

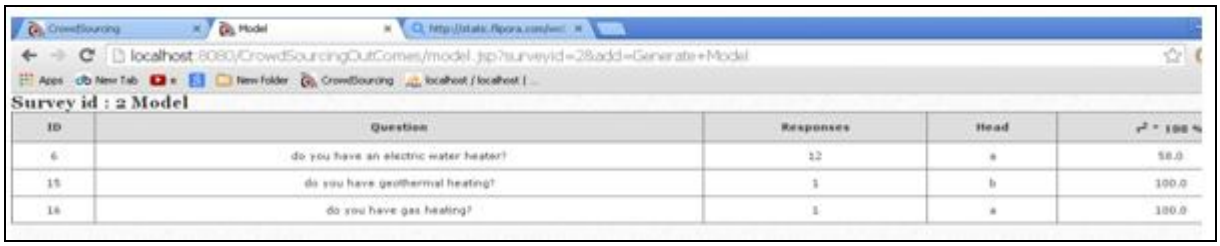


Figure 3: Model generated for Energy consumption (Energy)

Following figures show the result analysis.

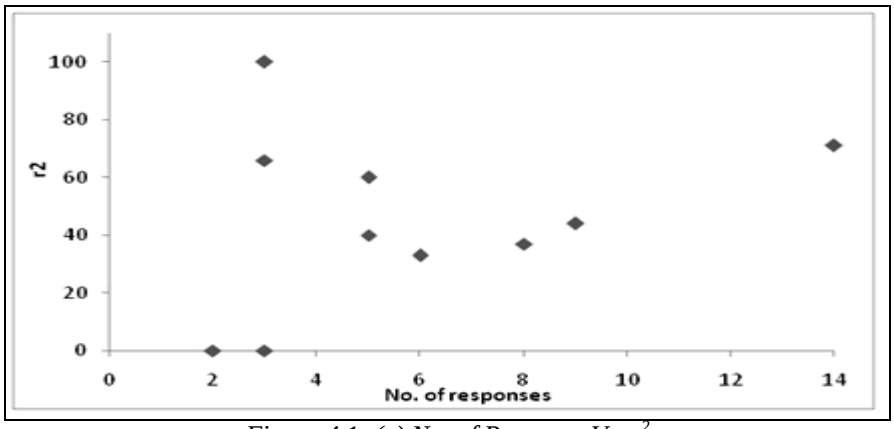


Figure 4.1: (a) No. of Response Vs. r^2

Above figure 4.1 (a) shows the relative predictive power of the 10 questions. Above result shows that most highly correlated factors (question ID 15,16,23) should be posed after initial 3 questions, since question no. 15,16,23 are dependent on question ID 3. And a weak correlation between the response rate and the r^2 values, indicating that more answers to questions would have likely produced improved results.

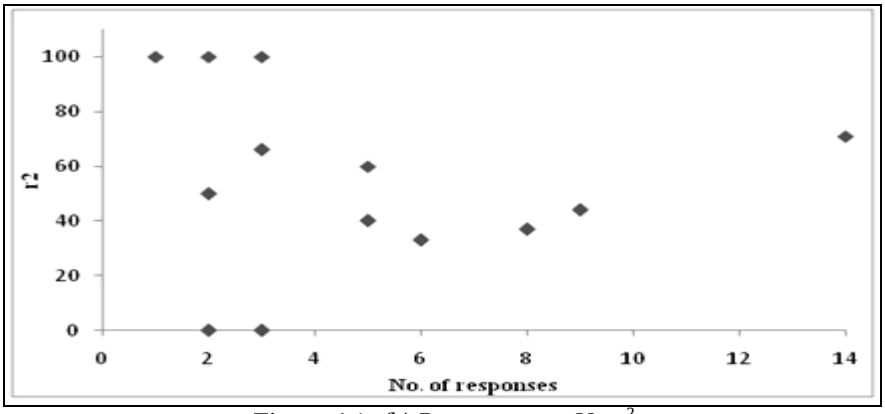


Figure 4.1: (b) Response rate Vs. r^2

Figure 4.1 (b) shows the relative predictive power of the 13 questions, indicating that more answers to questions would have likely produced improved results

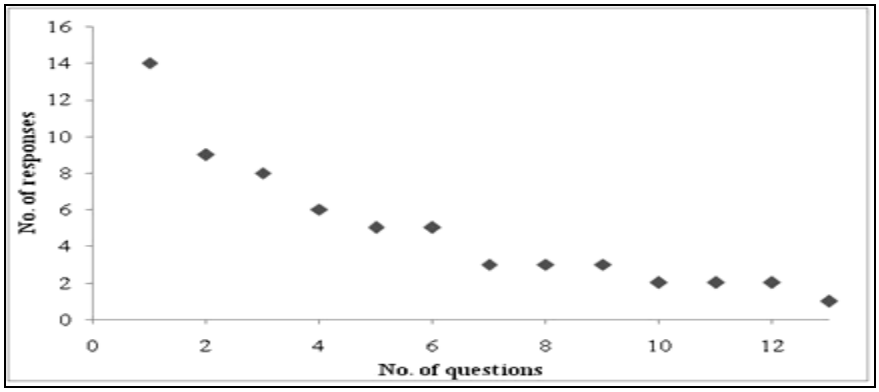


Figure 4.1 (c): No. of questions Vs. No. of responses

Above figure 4.1 (c) shows the number of responses for each question, sorted by the number of responses to facilitate the comparison. In instantiation of this concept, a web-based social network is developed to model residential electric energy consumption. Because of policy efforts to increase energy efficiency, many are working to provide consumers with better information about their energy consumption. Research on consumer perception of energy efficiency indicates that electricity customers often misjudge the relative importance of various activities and devices to reducing energy consumption. To provide customers with better information, numerous expert driven web-based tools have been deployed. In some cases these tools use social pressure as a means of improving energy efficiency , however the feedback provided to customers typically comes from a central authority (top-down feedback) and research on risk perception indicates that the public is often skeptical of expert opinions.

As discussed in above figure, the relative predictive power of the 10 questions. The results show that the most highly correlated factors were posed after the initial two seed questions (Fig. 4.1 a) and a weak correlation between the response rate and the r^2 values, indicating that more answers to questions would have likely produced improved results. Panels (c) and (d) show the distributions of r^2 values and the number of responses, to facilitate comparison with the BMI.

4. Discussion and Conclusion

The results are discussed below and some conclusions can be written down.

- The half of the participants provided energy data was most likely due to the effort associated with finding one or more electricity bills and entering data into the site.
- It is found that participants were reluctant or unable to provide accurate outcome data due to the challenge of finding one's electric bills.
- The result shows that most highly correlated factors (question id 15,16,23) should be posed after initial 3 questions, since question no. 15,16,23 are dependent on question id 3.
- It is observed that there is a weak correlation between the response rate and the r^2 values.
- This result is not consistent with the fact that owning an electric hot water heater increases electricity consumption. It appears either that this correlation was due to chance, or that ownership of a gas hot water heater correlates to some other factor, such as (for example) home ownership is correlated with electric water heater.
- The low response rate (unanswered questions) emphasized that the utility of this approach depends highly on ease with which user can access outcome data.
- It is possible to produce user-generated questions and answers, and that a trial with a larger sample size might provide more valuable insight.
- Finally, questions that were posed early in the trial gained a higher response rate, largely because many users did not return to the site after one or two visits. This emphasizes the importance of attracting users back to the site to answer questions in order to produce a statistically useful model.

5. References

1. J. Evans and A. Rzhetsky, "Machine science," Science, vol. 329, no. 5990, p. 399, 2010.
2. J. Giles, "Internet encyclopedias go head to head," Nature, vol. 438, no. 15, pp. 900–901, 2005.
3. D. C. Brabham, "Crowdsourcing as a model for problem solving," Convergence, vol. 14, pp. 75–90, 2008.
4. A. Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008.
5. M. Marge, S. Banerjee, and A. Rudnicky, "Using the amazon mechanical turk for transcription of spoken language," in Proc. IEEE International Conference on Acoustics Speech and Signal Processing, 2010.
6. N. Kong, J. Heer, and M. Agrawala, "Perceptual guidelines for creating rectangular treemaps," IEEE Transactions on Visualization and Computer Graphics, vol. 16, no. 6, 2010.

7. A. Kittur, E. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in Proc.Twenty-sixth annual SIGCHI conference on human factors in computing systems, 2008.
8. A. Kittur, "Crowdsourcing, collaboration and creativity," XRDS, vol. 17, no. 2, pp. 22–26, 2010.
9. Josh C. Bongard, Paul D. H. Hines,Dylan Conger, Peter Hurd, and Zhenyu Lu, "Crowdsourcing Predictors of Behavioral Outcomes", IEEE Transactions on Systems, Man, and Cybernetics. Updated on March 8, 2012