International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Predictive Models for Behavioral Outcomes through Crowdsourcing

Veena M.E

Computer Engineering Department, Pune University, Pune, Maharashtra, India

Abstract: ***To develop models from large datasets and determine which subset of data to mine is now becoming automated. Choosing what kind of data to be collected and place it first, requires an immediate apprehension by human which is supplied by domain expert. A new approach to machine science which demonstrates that even non domain experts can express clearly and precisely the features and provide values for those features so that they are predictive of some behavioural outcome of interest. It is done by creating a web platform where interaction between groups of people by responding to questions which help to predict behavioural outcome, results in dynamically growing online survey. This leads to predict the behavioral consequences of user with the help of their responses to survey questions formed.***

Keywords: Machine science, Crowdsourcing, Predictive Modeling, Human behavioral outcomes.

1. Introduction

To develop predictive model which map between set of outcomes and predictor variables give rise to many problems. While the model structure are specified with the set of predictive covariates statistical tools provides mature methods to compute model parameter. Team of experts are needed in such problems for each individual domain which results in excess loss of human efforts. For example, to choose appropriate questions related to respected domain an expert survey designer must be needed for that domain. The performance will increase in such a way that an engineer must keep correlation and suitable approach of design in order to judge which concept will be more efficient. Necessity of domain expert is the main drawback of this approach.

To understand the difficult problems will harness the effectiveness of result, however is done through using the knowledge of crowd. Modeling can be tested by an alternative way that is online crowd, which can be used to define potentially predictive variables to study by asking and getting response to question, so that a predictive model is developed.

A. Machine Science

Machine science consists of automation of many scientific concepts. But in case of machine science it is difficult to decide which variable of subset is to be selected. In addition, it is also very difficult to decide which variable to be automated. The prediction problem in machine science sometimes is unable to select the variable which can predict the outcome of interest.

To map between a set of predictor variables and an outcome, there are many problems in which one seeks to develop predictive models. The selection of variables for which data to be collected for evaluate hypothesis is one aspect of the scientific method that has not yet yielded to automation. To select the independent variables that might predict an outcome of interest and for which data collection is required, is however the prediction problem in machine science [1]. To test an alternative approach, modelling the experience and knowledge of crowds is used to propose, which potentially predictive variables to study by asking questions and provide the data by responding to those questions, is the goal of this approach. This paper introduces the method in which there is a motivation for non domain experts to formulate independent variables but also populate enough of variables to form successful modelling. This can be explained as follows. Users visit a site based on behavioural outcomes (The behavioural outcomes could be a body mass index or daily electricity consumption) is to be modelled. The user will provide their own outcomes (like their own consumption of electricity) and answer the questions that may be predictive of that outcome (like how much electricity they use daily). By ordinarily, different models are constructed in oppose to growing data sets predicting user's behavioural outcome. User can also post their own questions that, which becomes new independent variables when answer by other users in the modeling process. Thus to discover and populate independent variables will be done by user community [2].

B. Crowdsourcing

The method of Crowdsourcing was introduced by Jeff Howe. Crowdsourcing can be defined as "an act of outsourcing task, traditionally performed by employee or contractor which are now performed by large group of people". The Crowdsourcing research is involved in variety of fields such as computer science, management, and many other domains which have discovered Crowdsourcing as a useful approach.

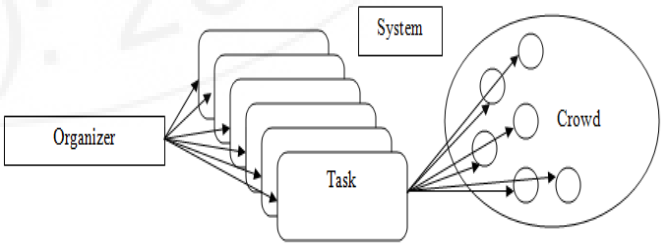

Figure 1: General Architecture of Crowdsourcing

Page Layout Above figure shows the general architecture of crowdsourcing, it shows that tasks are the works that are

assigned to crowd. Crowd is the group of people who performs those tasks. Organizer can be either organization or individual who assigns the tasks to crowd and evaluates the results. System facilitates organizer to assign the tasks to crowd and to review the results. In many cases, when some problems (like extra BMI or excess electricity consumption) occur, it is necessary to consult domain experts for solution of the problem. However this problem can be solved with the help of crowdsourcing i.e. user can solve their own problems by their own selves. The effectiveness of crowdsourcing is proved by the best example, Amazon's Mechanical Turk [3]. In this, it is explained that a "Human Intelligence Task" such as characterization of data, transcribing spoken language, or creating data visualizations can be accomplished with the help of group of people which is very difficult for a computer to accomplish alone.

2. Methodology

Application of crowdsourcing in cyber infrastructure is described in the paper [4], such that the following conditions are true.

1) Some human behavioral outcomes that are modeled are defined by the Investigator.
2) Investigator defines the model in this phase of system flow for which, investigator wants to predict the consequence then as per requirement can accept or pose a question. After analyzing the question if it is suitable for the module, whatever the accepted question from user the investigator can add it.
3) From user i.e. from human volunteers data required for our model is collected. They may or may not be domain expert.
4) The models are then continuously generated with automatic approach.
5) To propose new independent variables, volunteers are also motivated.

The overview of the system is as shown in below fig 2. This gives a clear illustration of how the investigator model, user model, and modeling engine work together and how they are related to each other to produce predictive models of the behavioral outcome of interest.

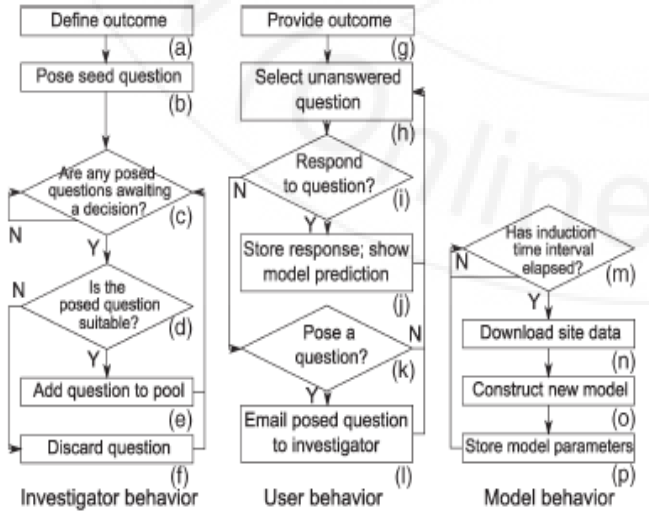

Figure 2: Overview of the system

Basically the system is divided into group of tasks, like first group of task is that it includes the working of the investigator model. User model and modeling engine, this group work together and produces the predictive model of the outcome of interest which is the second group of task. First of all the investigator creates a web site which define the human behavioral outcome which is to be modeled. The paper deals with the financial and health outcome which are investigated. First outcome of interest is the monthly electric energy conservation and second one is the body mass index. The site is initialized with the set of some questions which are related to the outcome of interest by the investigator. For example in case of consumption of fast food and obesity, we seeded the BMI website with the question "How often a week do you eat fast food?"

The users who visit the site individually give the values for the outcome of interest such as their own BMI (Body mass Index). The user then responds to the questions display on the site. The dataset is used to store the information provided by the user, in addition it is forwarded to the modeling engine. After receiving the data or information from the user, modeling engines begins working and construct a single matrix $\mathbf{A€}\ \mathbf{R}^{n\times k}$ and outcome vector **b** having length *n* from the collective responses of n users to k questions. **A** is combination of a_{ij} elements indicating the *i* user's response to question *j*, and each element $\mathbf{b}_i$ in **b** is the *i* users outcome of interest. Model of outcome was achieved by using the linear regression model.

The output of the modeling process is the vector **c** of length k+1 consists of model parameters. It also gives the output vector **d** having k length which stores the predictive power of each question: $\mathbf{d}_j$ stores the r^2 values obtained by regression only on column *j* of **A** against the response vector **b**. The two outputs vector **c** and vector **d** are stored in the data store. User can pose a question at any time and their own choice and design. These choice and design includes a Yes or No response questions, five level likert rating, or a number. Users were not forced to what kind of questions to pose. Once the user poses a question the suitability is checked by the investigator. The questions were considered unsuitable if it holds the following conditions: (1) the question revealed the identity of its author, thereby conflicting the Institutional Review Board approval for these experiments; (2) the question contained profanity or hateful text; (3) the question was inappropriately correlated with the outcome. If the question was specifically suitable it was added to the pool of questions available on the site; otherwise the question was discarded.

Each time a user visit to the site, they start with a new set of question, which are unanswered as well as additional data so that, interest in the site, their participation in the experiment should be maintained. After answering all the available questions, it shows a list of the questions, their answers, and some circumstances relevant information to indicate how their responses compared to those of others. The most important information shown to each user after responding to each question was the value of their actual outcome as they entered it (b_i) as well as their outcome as predicted by the current model ($\hat{b}_i$). After each response from a user

$$\hat{b}_i = c_0 + c_1 a_{i1} + c_2 a_{i2} + \ldots\ldots + c_k a_{ik} + \epsilon_i$$

If user i has not yet answered the question j then a_{ij} = 0, otherwise a_{ij} is set to the user's response.

The goal of this approach is to test an alternative approach to modelling in which the wisdom of crowds is controlled and make use to both propose which potentially predictive variables to study by asking questions and to provide the data by responding to those questions. Instead of using the linear regression approach the hierarchical regression approach is applied in the proposed method. We can increase the processing system and result analysis of the system, by applying n number of models instead of using single model as used in the existing system as the main model [5].

3. Benefits and Challenges

The participants play a vital role to highlight at least one behavioral consequence in both the cases like body mass index and daily electricity consumption. If the number of users will provide with the number of questions at a time to the website then the system will get overflow it is the challenge of this system. With the help of dynamic filtering of questions this problem can be overcome [4]. Other challenges includes,

A. User Fatigue

The system is the user fatigue, may happen that the user answers only a small instance of all questions and as a result, some question may get more response than others. As questions are added to question pool as per the users suggestions, so questions that are present at first will get the more response than others. The user may answer the questions that are less predictive than those which are more predictive and it leads to wrong prediction.

B. User Motivation

This is another challenge in this case we have to motivate the user each time to answer all the questions in the question pool to generate proper outcome result. If the user will not answer the question whatever the consequences happen that will not provide the proper outcome result. Depends on user's response, each time we have to motivate the user.

C. Rare Outcomes

Sometimes it may happen that the user suffering from the rare disease visit the website and gives the answers to questions it sometime may lead to wrong predictions of outcomes. All the problems we have discussed above can be overcome by motivating users to give proper answers and also to attempt all the questions.

D. Faster Result

As the system is semiautomatic means can generate the result in less time and less efforts. Whatever the time that required communicating with domain expert get saved as the system can generate the result without the help of domain expert.

4. Conclusions

A new approach to social science modeling has introduced in this paper, in which human behavioral outcome is generated by motivating the participants. User visits the web site and answering to the questions which wants to and not wants to be leads to hectic for the participant. So by applying distinctive approach using the different regression model, question ordering could be made systematically so that the user could not face the questions that he don't want to answer. Also rather than using of the single model the working of the system could be enhance by applying the n number of models. In future instantiations of the method would be to dynamically filter the number of questions that a user may respond to: As the number of questions approaches the number of users, this filter would be strengthened such that a new user is only exposed on a small subset of the possible questions.

References

[1] Josh C. Bongard, *Member, IEEE*, Paul D. H. Hines, *Member, IEEE*, Dylan Conger, Peter Hurd, and Zhenyu Lu, "Crowdsourcing Predictors of Behavioral Outcomes", IEEE Transactions on Systems, Man, and Cybernetics: systems, vol. 43, no. 1, January 2013.
[2] Josh C. Bongard, "Crowdsourcing Predictors of Behavioral Outcomes," IEEE transactions on knowledge and data engineering, 2013.
[3] A. Sorokin and D. Forsyth, "Utility data annotation with Amazon Mechanical Turk," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Anchorage, AK, 2008, pp. 1–8
[4] Barde, G. S., Dhavane, R. S., Khole, P. K., Mehta, V. P., &Dasgupta, A. Novel Semiautomatic Crowdsourcing Predictors for Faster Statistical Analysis Based Upon User Inputs.
[5] D. Wightman, "Crowdsourcing human-based computation," in Proc. 6th Nordic Conf. Human-Comput. Interact.- Extending Boundaries, Reykjavik, Iceland, 2010.
[6] P Jadhav, S., and M. R Patil. "Performance Analysis for Crowdsourcing Context Submission using Hierarchical Clustering Algorithm and Classification". *International Journal of Computer Applications* 92.11 (2014): 33-37.
[7] Josh C. Bongard, Member, IEEE, Paul D. H. Hines, Member, IEEE, Dylan Conger, Peter Hurd, and Zhenyu Lu Crowdsourcing Predictors of Behavioral Outcomes IEEE Transactions on Systems, Man, and Cybernetics. Updated March 8, 2012
[8] Vinay V. Mandhare, Vinod Nayyar. "A Survey on Crowdsourcing and Behavioral Outcome" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, January 2014.

Author Profile

Veena M.E received the B.E. degree in Computer Science and Engineering from Jawaharlal Nehru Institute of Technology in 2007. Now Pursuing M.E degree in Computer Engineering from JSPMs Rajarshi Shahu College of Engineering.