ISSN: 0975-5662, September, 2014 www.ijrct.org

Predicting Outcomes Based on Hierarchical Regression

Srilatha Thota[1], M.Tech, Computer Science and Engineering, E mail: t.srilathab9@gmail.com
Vijaykumar Janga[2], Asst. Professor, Department of CSE, E mail: vijaykumar.janga@gmail.com
Fasi Fhmed parvez[3] Associate professor and HOD, Department of CSE,Email:parvez40509@gmail.com
#, **BALAJI INSTITUTE OF ENGINEERING AND SCIENCES**, Warangal, Telangana, India

Abstract:
Large datasets tends to developed models and determines which subset of data to mine is becoming automated. However, selecting the kind of data and place is in first place which requires human experience supplied by domain expert. This paper gives new approach to machine science demonstrating that non domain experts can collectively formulate features and provide values for those features so that they are predictive of some behavioural outcome interest. This was accomplishing by web platform where group of people get interact with each other by responding to question which help to predict behavioural outcome. Which result in dynamically growing online survey.

1.0 Introduction

To develop predictive model one seeks many problems, which map between set of outcomes and predictor variables. The researchers are still providing new tools for inferring the structural form of non-linear predictive models, given good input and output data [1]. However, choosing of task which will potentially predictive variables to study is largely a qualitative task that requires substantial domain expertise. The need for the involvement of domain experts can become a bottleneck to new insights. However, if the wisdom of crowds could be harnessed to produce insight into difficult problems, one might see exponential rises in the discovery of the causal factors of behavioral outcomes, mirroring the exponential growth on other online collaborative communities. While the model structure are pre-specified with set of predictive variables, statistical tools provides mature methods to compute model parameter. The main goal of the research is to test an alternative way to modeling in which online crowd can be used to define potentially predictive variables to study by asking and getting response to question, so that an predictive model is developed.

1.1 Machine science

Machine science is a new scientific technology which very interesting to find, analyse, classify the data to generate hypothesis and develop models [1] Machine Science is a hot topic in the theory and philosophy of modern science, with recent claims that "within a decade, even more powerful tools will enable automated, high-volume hypothesis generation to guide high-throughput experiments in biomedicine, chemistry, physics, and even the social sciences" [1]. This paper introduces a method in which non domain experts can be motivated to formulate independent variables but also populate enough of variables to form successful modelling. This can be well explained as follows. Users or user 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 (Such as 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.

1.2 Crowdsourcing

Now days there is a rapid growth in user generated contend on the internet is one of the best example that how the user interactions can effectively solve the problem under the explicit management by team of experts **Crowdsourcing** is the process of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, and especially from an online community, rather than from traditional employees or suppliers. This process is often used to subdivide tedious work or to fund-raise startup companies and charities, and can also occur offline. It combines the efforts of numerous self-identified volunteers or part-time workers, where each contributor of their own initiative adds a small portion to the greater result. The Crowdsourcing has been used in number of research and commercial applications. For an example the Wikipedia in which set of information is being published and later on number of people can add and enhanced the information finally the best of the best enhanced data is available for the user which cannot be done by using a single computer alone and could be expensive to achieve through expert domain process.

"Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken". Wikipedia is the best example that explains how the online collaborative approach can solve the difficult problem very easily without any expenses. This survey paper reports on task with direct motivation: one could be a house hold energy usage task in which user are able to understand their own energy consumptions which helps to improve their energy saving; and another could be the body mass indexing task which helps user to understand their lifestyle choice so that they can live a healthy life .These both instantiations include comparative approaches in which participants compare

with each other and also allow to predict quality of question that participant provide.

2.0 Existing System:

When considering the non-ideal case of a given system, nonlinear dynamics are usually what appears between the linear elements of the ideal model. The system's inputs and some of its states may be physically limited depending on the system characteristics. Generally, under nonlinear dynamics, a finite-time optimal control sequence can be obtained more easily than an infinite-time optimal controller. Since the MPC controls are calculated repeatedly over the finite horizon, it is reasonable to believe that it would be easier to use MPC than the infinite-time optimal controls. The optimization problems solvers over a finite horizon can be applied to a wide variety of systems, including nonlinear systems and time-delay systems. Thus, MPC has a large population of potential applications even for nonlinear systems [4]. According to [5], the key characteristics and properties of nonlinear model predictive control (NMPC) are:

• NMPC allows the direct use of nonlinear models for prediction.

• NMPC allows the explicit consideration of state and input constraints.

• In NMPC a specified time domain performance criteria is minimized on-line.

• In NMPC the predicted behavior is in general different from the closed-loop behavior.

• For the application of NMPC typically a real-time solution of an open-loop optimal control problem is necessary.

• To perform the prediction the system states must be measured or estimated.

Let us give the mathematical formulation of a general NMPC problem. We consider the following discrete timeinvariant nonlinear system:

C2 is a nonlinear function describing the system dynamics, and f(0, 0) = 0. If U = R^m and X = R^n, the system x(i + 1) = f(x(i), u(i)) is defined as an "unconstrained system". In the case of ui U Rm or xi X Rn, the system x(i + 1) = f(x(i), u(i)) with constraints u(i) U R^m and x(i) X 0 R^n is called a "constrained system." In the case of constrained systems, U is usually taken as a compact set including the origin.

3.0 Proposed System

However, all approaches have some drawbacks, which encourage us to try a fourth approach – hierarchical regression. The hierarchical regression model has already been successfully applied in other sub-disciplines of sociology such as family planning (e.g., Entwisle et al. 1984; Hirschman & Guest 1990) and education studies (e.g., Anguiano 2004), but it has not yet been tested in migration studies. In this article, we demonstrate how to apply the hierarchical regression model in migration decision making. We ask the question: Can migration studies be improved by using a multi-level approach that includes a mix of individual- and aggregate-level demographic, socio-economic, and biogeophysical factors? In the following sections, we first critique the mover-stayer approach, the multivariate regression approach, and the combination of the two, as well as the methodological advantages of the hierarchical regression approach. Second, independent variables (determinants of migration) used in this analysis are reviewed from the perspective of migration

decisionmaking. Third, we analyze the data on migration and the independent variables at two levels by formally specifying a hierarchical regression model. Finally, findings are summarized regarding the advantages of the hierarchical regression model. In this context, limitations and further studies are suggested. The hierarchical regression model, which already has been applied to family planning decisions (e.g., Entwisle et al. 1984; Hirschman & Guest 1990) but has not been explored in migration decision studies, has the potential to solve all the problems mentioned above. In demography, data are often structured hierarchically: Public Use Microdata Sample (PUMS) files are familiar examples which describe individual characteristics, household characteristics and housing unit characteristics for geographic Public Use Microdata Areas (PUMAs). Summary files provide aggregated attributes for areal data at block, partial block group, block group, census tract, county and state levels. Because of the characteristics of hierarchies, current studies that focus only on one level of the variables can only explain variations at that level (Bryk & Raudenbush 1992). This limitation has generated concerns of ecological or atomistic fallacies (Green & Flowerdew 1996; Robinson 1950; Voss et al. 2004; Wrigley et al. 1996). The advantages of the hierarchical regression models over the traditional mover-stayer model and the standard multivariate regression approach can be summarized as follows. First, because the hierarchical regression model can include spatial analysis when one of the hierarchical levels is geographic, it inherits advantages from spatial econometrics to account for the geographic heterogeneity. Second, the variations across groups can be estimated easily in a hierarchical regression model. For example, Tunali (2000) has applied detailed econometric models to study the move/stay decision using microdata in Turkey, and he built many models to examine and compare the effects of heterogeneous variables. However, his analysis could be easily handled by a hierarchical regression approach, where the effects of heterogeneous variables can be nested in the hierarchical models. Third, because the variations within- and acrossgroups can be estimated, the reliability of the coefficients (i.e., the ability of independent variables to explain the strength of relationships with moving probabilities) can be estimated. Fourth, the hierarchical regression approach combines both individual characteristics and aggregate-level characteristics in a model, allowing us therefore to avoid both ecological and atomistic fallacies in interpretation of analysis results (Robinson 1950).

4.0 Conclusion:

Based on this analysis, three advantages of hierarchical regression for studying migration are summarized. First, hierarchical regression can easily integrate heterogeneous variables at the aggregate level into one model, and their significances can be estimated. Second, the coefficient reliability of Level 1 variables can be estimated based on within- and across-group variance, and can then be used to re-estimate the coefficients of Level 1 variables. Third, the way that the hierarchical regression combines both individual and aggregate characteristics avoids the debates of ecological and atomistic fallacies. This is very important because individual behaviors are assumed to be influenced by, and their

aggregation is assumed to influence, the characteristics of the residential area. While the individual and areal linkage of migration studies is crucial for housing policy-making, it has long been ignored

5.0 References:

[1] J. Bongard and H. Lipson, "Automated reverse engineering of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 104, no. 24, pp. 9943–9948, 2007.

[2] J. Evans and A. Rzhetsky, "Machine science," *Science*, vol. 329, no. 5990, p. 399, 2010.

[3] R. D. King, K. E. Whelan, F. M. Jones, P. G. K. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell, and S. G. Oliver, "Functional genomic hypothesis generation and experimentation by a robot scientist," *Nature*, vol. 427, pp. 247–252, 2004.

[4] R. King, J. Rowland, S. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. Soldatova *et al.*, "The automation of science," *Science*, vol. 324, no. 5923, p. 85, 2009.

[5] J. Bongard, V. Zykov, and H. Lipson, "Resilient machines through continuous self-modeling," *Science*, vol. 314, pp. 1118–1121, 2006.

[6] J. Giles, "Internet encyclopedias go head to head," *Nature*, vol. 438, no. 15, pp. 900–901, 2005.

[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 KNOWLEDGE AND DATA ENGINEERING YEAR 2013

[8] Vinay V. Mandhare* Vinod Nayyar " A Survey on Crowdsourcing and Behavioral Outcome", ijarcsse Volume 4, Issue 1, January 2014

[9] Christophe Tricaud and YangQuan Chen "Linear and Nonlinear Model Predictive Control Using A General Purpose Optimal Control Problem Solver RIOTS 95", 2008 IEEE.

THOTA SRILATHA, M.Tech (COMPUTER SCEINCE AND ENGINEERING) in BALAJI INSTITUTE OF ENGINEERING AND SCIENCES, NARSAMPET. Area of interest is DATAMINING, NETWORK SECURITY.

Vijaykumar Janga, M.Tech(PhD) working as Assistant Professor in BIES, and a Research Scholar in JNTU-Hyd.
Area of interest is Data mining, Information retrieval.

Fasi Ahmed Parvez working as Associate professor and HOD BALAJI INSTITUTE OF ENGINEERING SCIENCES-NARSAMPET, with 12+ years of Experience. Completed M.Tech from JNTU Hyderabad in2010.

ISSN: 0975-5662, September, 2014 www.ijrct.org

SUBJECT INTERESTS are programming languages, database management system and data ware house & data mining.