ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

Predictive Modeling of Behavioral Outcome through Non Domain Expert Method Using Naive-Bayes Technique

Ms. Veena M. E[1], Mr. Sanjay B. Thakare[2]

1. M. E. Second Year Student Department of Computer Engineering JSPM's Rajarshi Shahu College of Engineering, Tathawade Savitribai Phule Pune University, India
2. Associate Professor Department of Computer Engineering JSPM's Rajarshi Shahu College of Engineering, Tathawade Savitribai Phule Pune University, India

Manuscript Info

Manuscript History:

Received: 15 July 2015
Final Accepted: 22 August 2015
Published Online: September 2015

Key words:

Predictive Modeling, Naive Bayes Method, Human Based Computation, Behavioral Outcome

Corresponding Author

Ms. Veena M. E

Abstract

Predictive models validate the anticipation of certain aspects of human behavior, such as goals, actions and preferences. In order to develop predictive models, there are many approaches in machine science which describes the computational techniques to develop models. This paper explains an approach that the non-domain experts can collectively design an attributes such that which subset of data to study and provide values for those attributes which helps to predict some behavioral outcome of interest. This is organized by allowing human groups to interact both by answering the questions and also pose questions their peers through web platform, likely to help in predictive modeling. However in literate linear regression technique is used to predictive modeling of behavioral outcomes. Regression creates predictive models with numerical or continuous target attributes. If the target attribute contains continuous values, it creates a regression problem that cannot explicitly refer to all target categories that are used in the model. The major challenge with this model is it over fit the data. This paper presents an improved method that uses Naive Bayes for modeling the prediction. Naïve Bayes overcomes the problems in earlier technique by dealing with discrete or categorical target attributes. Experimentally it is observed that the accuracy of Naive Bayes approach for behavioral prediction is better than the linear regression.

INTRODUCTION

To map between set of outcomes and predictor variables while developing predictive modeling gives rise to many difficulties. Choosing potential predictive variables makes the predictive modeling efficient. However, need of domain expert in qualitative tasks of selecting which predictive variables for which to collect data in the first place. For example, a survey designer and domain expertise must work together to determine questions that may identify predictive co variables. This helps determine which variables can be systematically modified in order to obtain effective performance. The necessity for domain expert involvement can cause a constriction to new situations. These constrictions can be overcome by making use of the experience and knowledge of crowd. Controlled use of the experience and knowledge of crowd may cause the exponential rise in the discovery of the casual factors of behavioral outcomes. Thus to choose an alternative approach to modeling and potentially predictive variables, sagacity of the crowd is used interrogating by questions and to provide data by responding to those questions. This human based computation, which is a non domain expert method considered as an effective method in predictive modeling of behavioral outcome of interest. The main objective of this paper is to develop a Predictive modeling

system, which enables non-domain experts to collectively frame the characteristic and provide values for those characteristic such that they are the predictors of outcome of interest. The approach used in this system is Naive-bayes method, the probabilistic model of naive Bayes method is based on Bayes theorem.

A. HUMAN BASED COMPUTATIONS

Human based computation is interchangeably called as Crowdsourcing. It is the process of collecting deserved services, ideas, or content by appealing services from a large group of people, and especially from an online community, preferably than from traditional employees or suppliers [1]. Human based computation is often used to divide tedious work and has occurred successfully 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. Human based computations are also used for data collection. The Amazon Mechanical Turk (MTurk) is a well known crowd sourcing Internet marketplace that enables individuals and businesses to coordinate the use of human intelligence to perform tasks that computers are currently unable to do. The Crowdsourcing research is involved in variety of fields such as computer science and informatics, management, cyber security and many other domains which have discovered human based computation as a useful approach.

B. NAIVE BAYES METHOD

Naive Bayes method is a statistical method for classification as well as a supervised learning algorithm. The Naïve Bayes classifier is based on Bayes theorem with independence assumptions between predictors. In Naive Bayesian model, there is no complication in iterative parameter estimation which makes it particularly useful for very large data sets. Naive Bayes provide rapid model constructing and scoring for relatively small size of data [6]. It is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. A naive Bayes classifier has two types of variables: the class variable C and a set of predictive features $x = \{x_1, x_2, x_3, \ldots, x_n\}$. Below figure represents the naïve bayes structure.

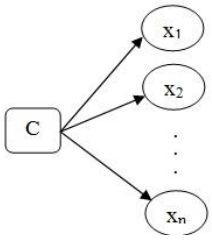

Fig. 1. Naïve Bayes Structure

In Naive Bayes, the data that was used while constructing the model is also used for verification and testing model accuracy. A Bayesian model does not separate the data as one portion to construct models and testing it on another portion. This paper is composed further as: Section II describes the motivation and challenges of the system. Section III talks about related work on the predictive modeling technique and crowd sourcing. Section IV presents implementation details, algorithms used, mathematical model. Section V depicts results. Section VI draws conclusions and presents future work.

An approach in machine science which demonstrates that non-domain experts can collectively formulate attributes and provide values for those attributes such that they are predictive of some behavioral outcome of interest. The human groups interact to both respond to questions likely to help predict a behavioral outcome and pose new questions to their peers. This motivated us to allow human groups to interact with each other, which helps to choose which potential variable to study. This result in a dynamically growing online survey, but the result of this cooperative behavior of crowd also leads to models that can predict the users outcomes based on their responses to the user-generated survey questions. The participants play a vital role to highlight behavioral consequences. 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. The other challenge of the system is the user exhaustion may happen that the user answers only a small instance of all questions and due to this some question may get more response than others. As we know that the questions get

added to question pool as per the user suggests it. 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.

II. RELATED WORK

Aniket Kittur, Ed H. Chi, Bongwon Suh [9], explains how to manage the effort and experience of crowd and this crowdsourcing has been used effectively in number of commercial and research applications. The best example of how crowdsourcing can be useful, consider Amazons Mechanical Turk, which is a crowdsourcing internet market place. In this crowdsourcing tool, a human explains a human intelligence event such as distinguishing data, arranging for spoken language, or creating data visualizations. The tasks that are difficult to manage with computers alone, however is possible by involving large group of people in different locations and it costs much to complete successfully through traditional expert driven processes.

Paul Clough, Mark Sanderson, Jiayu Tang, Tim Gollins, Amy Warner [2], this paper presents comparing similar assessments gathered using crowdsourcing with the involvement of domain expert to evaluate different search engines in a large government archive. But there are some limits: correlations between the crowsourced workers and the expert assessor were lower for certain kinds of queries.

Neelamadhab Padhy and Rasmita Panigrahi [5], proposed a prediction for the workers in the PR Department of Orissa, linear regression technique is used. In this technique, all attributes considered are numeric and linear regression technique is applied for prediction. This proposed work experience a disadvantage of linearity. In some cases that uses median regression technique for prediction, due to the data discloses the non linear dependency then the best results may not be obtained and also it has the high computational cost. So, this paper also suggests using objective data and formal regression models, which uses the mean computation. This is the simplest version among various regression modules is linear regression model. Linear Regression technique which takes the lesser time as compared to Least Median Square Regression.

Dursun Delen, Glenn Walker, Amit Kadam [8], this paper presents the relative study of multiple prediction models for cancer survivability shows the relative prediction ability of different data mining methods. The cooperative results indicated that the decision tree induction method produced the best with a classification accuracy of 93.6 % which is better than the ANN model shows the second best with a classification accuracy of 91.2 %, and the logistic regression model shows the worst with a classification accuracy of 89.2 %.

Josh C. Bongard, Paul D. H. Hines, Dylan Conger, Peter Hurd, and Zhenyu Lu [12], this paper explains the Prediction of BMI of the participants, linear regression technique is used. The major challenge was that the number of questions became visible to the number of participants on the BMI Web site. This gives rise to the possibility that the models may have overfit data that is, random error or noise instead of the underlying relationship with the data. It occurs when a model is extremely complex, such as having excessive parameters relative to the number of observations. A model that has overfit generally does not have better predictive performance, as it can overestimate minor variations in the data. In order to avoid over fitting, it is necessary to use additional techniques like cross-validation, regularization, Bayesian priors on parameters or model comparison.

Guzmn Santaf, Jose A. Lozano, and Pedro Larraaga [7], describes the Bayesian model-averaging approach for an unsupervised naive Bayes classification model. This approach allows acquiring the parameters for the approximate model averaging cluster model. These parameters are acquired in the same time complexity necessitate to learn the maximum likelihood model. The model-averaging over selective naïve bayes structure is attained by the Expectation Model Averaging (EMA) algorithm.

Shaoyan Zhang, Christos Tjortjis, Xiaojun Zeng, Hong Qiao, John Keane [4], this paper proposes a comparison of logistic regression with six data mining techniques decision trees, association rules, Neural Networks, naive Bayes, Bayesian networks and Support Vector Machines for the prediction of overweight and obese children at 3 years. To compare accuracy: the prediction rates from logistic regression, decision tree and association rules are poor. The neural network performs better than the above mentioned algorithms, but not as well as the Bayesian algorithms and SVMs. Bayesian algorithms are profitable than the SVMs in terms of overall prediction rate as compared to SVMs sensitivity prediction rate.

III. IMPLEMENTATION DETAILS

A. SYSTEM ARCHITECTURE

The system described in Fig 2 illustrates system architecture to predict the human behavior modeling. The architecture consists of three modules User, Investigator and Model behavior. These three modules work together to generate a predictive model of the outcome of interest.

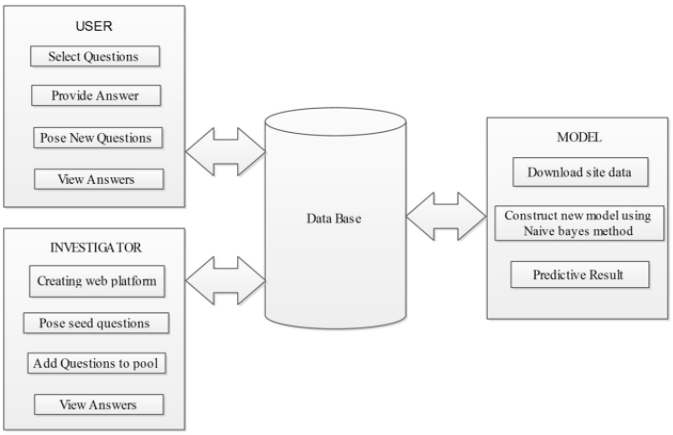

Fig. 2: System Architecture

Investigator is in charge for defining some behavior based outcome that is to be modeled. Investigator starts it by building a web platform and presents some set of beginning questions. Investigator is also responsible for setting question and answers frequency. User who visits the site should register to the site and can answer to the questions of their own interest and the answers are stored in data set. User may also present their own questions to site in addition to that he may also view answers and check polling. Investigator verifies the users questions. If the question is suitable for the particular context of behavior modeling then, investigator adds the question to pool and discards otherwise. In model behavior, we are using Naive-Bayes for prediction of outcome of interest. Naive-bayes classification algorithm is a supervised learning method and based on Bayesian theorem [13],

$$P(H|E) = (P(E|H) * P(H)) / P(E)$$

The basic idea of Bayess theorem is that the outcome of a hypothesis (H) can be predicted based on some evidences (E) that can be discovered. From Bayess theorem, we have a priori probability of H or P (H): This is the probability of an event before the evidences is discovered. A posterior probability of H or P(H|E): This is the probability of an event that occurs after the evidence is identified. The system input is divided as the number of answers to the respective number of questions. The answers are then compared with other answers and this leads to combined output as a prediction result. This behavior defines the divide and conquers strategy of the system. The system takes n inputs as answers to respective n number of questions by the users. These answers are stored in the common data set. Naive bayes predictive modeling technique is applied to these answers and single prediction result is obtained. Thus satisfying the multiplexor logic design of the system.

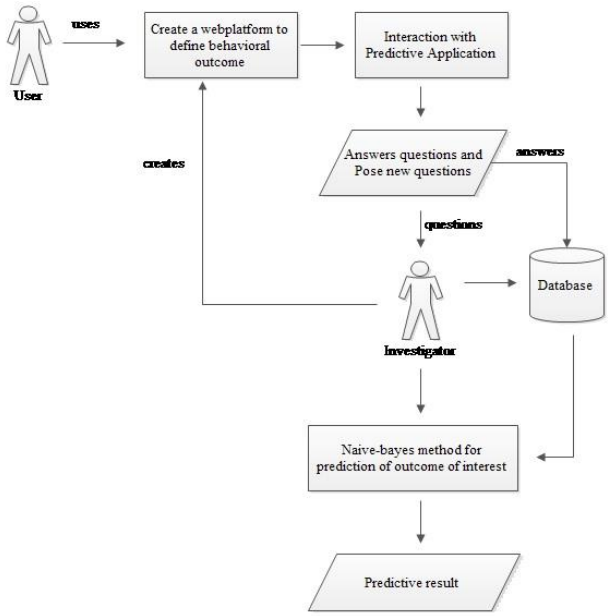

Fig. 3: Data Flow Diagram

B. ALGORITHM

In proposed system, we are using Naive-Bayes classifiers algorithm. The probabilistic model of Naive-Bayes classifier is based on Bayes theorem with the assumption that features in a dataset are mutually independent. The algorithm is as follows:

Input: Datasets and Question sets.

Output: Prediction and Probability of behavioral outcome of interest of user.

1: if user find interest in Dataset then

2: user = SelectedDataset;

3: userInfo[i] = user (ID, Name, Age, Gender);

4: userAnsQuestion[j] = question (ID, questions, Ans);

5: QuestionType ← Select Question and Answer into userAnsQuestion[j];

6: end if

7: while (Ans! = Null) do

8: PredictionSet ← view all user answer to a particular question and set Lower bound and Upper bound to prediction;

9: end while

10: NewUserSet ← Select those entire users whose values belong to Lower bound and Upper bound values.

11: ProbabilitySet ← Select users into NewUserSet with calculating their probabilities with Naive-Bayes Probability theorem.

12: return prediction to User and Investigator.

C.MATHEMATICAL MODEL

S is the system set for predictive modeling of behavioral outcome of interest. The system uses Naive-Bayes method this performs some computations on user's data and predicts the result of behavioral outcome of interest.

S= {I, O, f }

I →input set which contains,

I = Q, U, U_I , A

Q → is the Question Bank.

Q = {Q_1 , Q_2 , Q_3 , . . . , Q_n }, where n is the number of questions. U = {U_1 , U_2 , U_3 , . . . , U_k }, where k is the number of users.

U_I → is the user profiles.

U_I = {P_1 , P_2 , P_3 , . . . , P_k }

P_i is user profile where k is the number of users. A → matrix of k*n+1

O → output of the system

x → is a features or dependent variables in naive-bayes classifiers.

x={x_1 , x_2 , . . . , x_n } , where n is the number of features, and assigns probabilities to this instances is, $p(c_k | \{x_1 , x_2 , . . . , x_n \})$ for each k is possible outcomes or classes.

Conditional probability is given by, $p(c_k |x) = p(c_k * p(x|c_k)/p(x)$

if features are conditionally independent of each other, i.e i=j, is given by, $p(x_i |c_k , x_j) = p(x_i |c_k)$

thus the model can be expressed as, $p(c_k |x_1 , x_2 , . . . , x_n) \propto p(c_k , x_1 , . . . , x_n)$

$\propto p(c_k)$, $p(x_1 |c_k)$, $p(x_2 |c_k)$ . . .

$\propto p(c_k) p(x_i |c_k)$ where i=1 to n

f(x) is the system functionality which generates predictions using n number of questions having responses from k number of users, who are having different profiles. Predictions obtained are then manipulated by using naive- bayes algorithm to get probability of prediction results.

D. EXPERIMENTAL SETUP

The system has been developed using C#.Net programming language. A graphical user interface has been designed using ASP.Net to make it more interactive and user friendly. The system does not require any specific ardware to run; any standard machine is capable to run an application.

IV. EXPERIMENTAL RESULTS

We have taken car dataset from uci repository to predict the miles per gallon attribute. Dataset contains five attributes namely acceleration, displacement, cylinders, horse power and weight and these attributes can have either discrete or continuous values. The prediction is done for hundred instances; sample dataset is shown in the table II and table III.

TABLE I: Memorization Parameters

Notations	Meaning
P(H)	Probability of Hypothesis
P(E)	Probability of Evidences
P(H|E)	Posterior Probability
Q_i	Instances of Questions
U_I	User Profiles
P_i	Instances of User Profiles
x_i	instances of features
c_k	Instances of Probabilities to k possible classes

TABLE II: Car dataset to predict miles per gallon

Sample No.	Acceleration	Cylinder	Displacement	HorsePower	Weight
1.	11.5	8	350	165	3693
2.	11	8	318	150	3436
3.	12	8	304	150	3433
4.	10.5	8	302	140	3449
5.	10	8	429	198	4341

TABLE III: Users Predicted miles per gallon

Actual Outcome	Predicted Outcome
15	15.576
18	16.819
16	16.414
17	16.924
15	12.894

We have developed the website to predict the car's miles per gallon in which we are using the car datasets. The website consists of simple login page and four other interactive pages. After user's login to the site, home page

contains further instructions directing users to perform tasks on the website. In home page users can also has a choice to select one of the four predictions of behavioral outcome like monthly electricity bill, crime rate, car's miles per gallon and BMI.

Here we considered cars miles per gallon prediction. Pose Question page allows users to pose new questions and also mentioned the type of answers are expected for the respective question. These user posed questions are first sent to the website admins mail. Administer investigates the user question is suitable for the respective prediction, if it is valid the the question is added to the question pool otherwise discarded. The answer page allows users to answer the questions and provide them with the information related to the each answered questions. Answer review page allows users to review answers and provides users to see answers of other users and compare themselves with others. Result page displays the users predicted results of behavioral outcome of interest. The website also allows the users to provide their actual outcome and compares with the predicted outcome and displays the error estimation between the actual outcome and the predicted outcome.

In this experiment we considered car dataset to predict the cars miles per gallon which is having five attributes. Users enter the site and answers all the questions related to the prediction and user also provide the actual miles per gallon. After answering all the questions the user's prediction of the cars miles per gallon is displayed in the result page. The user's prediction results are also stored in the database and admin can view the results of all the users as shown in figure

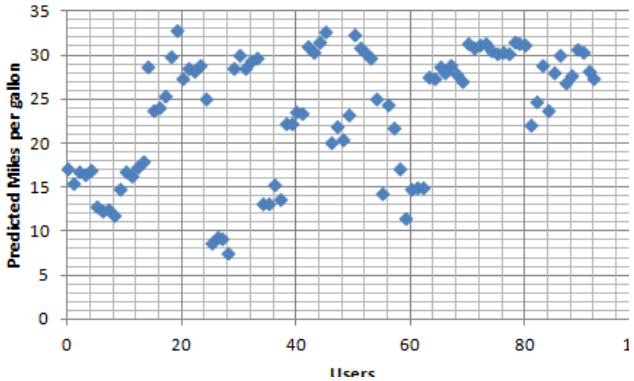

Fig. 4: Graph Showing predicted car milage per gallon

The prediction results obtained by the naive bayes technique, defines the correlation between the questions and the users answer. This results shows that there exists a power law relationship between the predictive questions. The success definition of the predictive modeling lies in choosing the potential predictive variables and providing the values for these variables.

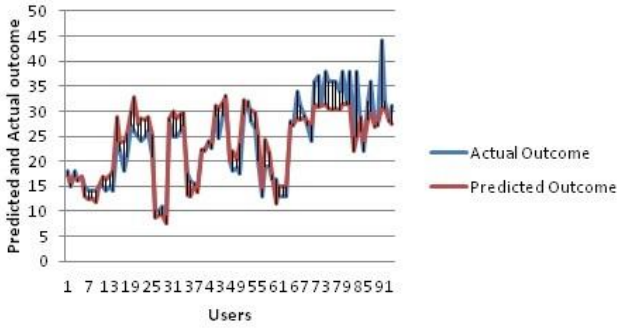

Fig. 5: Graph Showing Error Estimation between Actual and Predicted Outcome

The graph shows the results associated with the predicted answers and the samples in the dataset. These are obtained by the values associated with the attributes and these values are given by the user during the prediction of behavioral outcome of interest.

In literate Linear Regression technique is used for predictive modeling. In this work naive bayes classification is used for predicting the behavioral outcome of interest. Both methods train attributes weights W_j for the linear decision function j W_j X_j. The difference is how you fit the weights from training data. In Naive Bayes Classification, we can set each attributes weight independently, based on how much it correlates with the class label. By contrast, in linear regression, all the weights for the attributes are set together such that the linear decision function incline to be high for positive classes and low for negative classes.

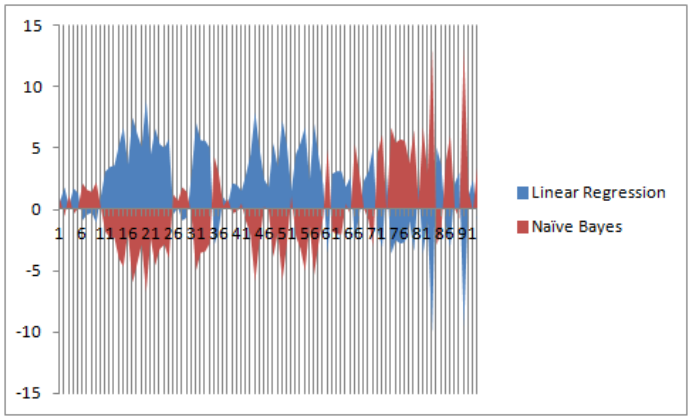

Fig. 6: Graph Showing Comparison between Linear Regression and Naïve Bayes Classification

In figure, we see how Naive Bayes and Linear Regression are compared over the same dataset. The plots are the error verses the size of the training data. The blue line is Linear Regression and the red line is Naive Bayes. This graph supports the claim that Naive Bayes works better over smaller training data as compared to linear regression. The graph also indicates that the error estimation of Naive bayes method is less compared to linear regression.

V. CONCLUSION AND FUTURE SCOPE

In this paper we proposed an approach for predictive modeling of behavioral outcome of interest by Naive-Bayes method. Naive-Bayes method is a statistical and supervised learning method for classification. It uses a small amount of training data to calculate the parameters that are necessary for classification. Because independent variables are accepted, only the variances of the variables for each class need to be determined and not the entire covariance matrix. In this work we have calculated the error estimation between the actual outcome and predicted outcome. One method to represent in future would be dynamically filtering the number of questions that a user may respond to: As the number of questions approaches the number of users, this filter would be build up such that a new user is only exposed on a small subset of the possible questions.

REFERENCES

1. D. Wightman, Crowdsourcing human-based computation, in Proc. 6th Nordic Conf. Human-Computer Interact. - Extending Boundaries, Reykjavik, Iceland, Copyright 2010 ACM ISBN: 978-1-60558-934-3.

2. Paul Clough, Mark Sanderson, Jiayu Tang, Tim Gollins, Amy Warner, Examining the Limits of Crowdsourcing for Relevance Assessment, Published by the IEEE Computer Society 1089-7801/13/ 2013 IEEE.

3. Jacob Abernethy, Rafael M. Frongillo, A Collaborative Mechanism for Crowdsourcing Prediction Problems, Division of Computer Science, University of California at Berkeley, ACM 1528- 2/10/1200 2010.

4. Shaoyan Zhang, Christos Tjortjis, Xiaojun Zeng, Hong Qiao, Iain Buchan, John Keane, Comparing data mining methods with logistic regression in childhood obesity prediction, Inf Syst Front 11:449460, Springer Science + Business Media, LLC 2009

5. Neelamadhab Padhy and Rasmita Panigrahi, Assistant. Professor and research scholar, Gandhi Institute of Engineering and Technology, GIET, Gunupur, Data Mining: A prediction Technique for the workers in the PR Department of Orissa, International Journal of Computer Science, Engineering and Information Technology, Vol.2, No.5, October 2012.

6. Marco A. Wiering, Hierarchical Mixtures of Naive Bayesian Classifiers, Intelligent Systems GroupInstitute of Information and Computing Sciences Utrecht University.

7. Guzmn Santaf, Jose A. Lozano, Member, IEEE, and Pedro Larraaga, Bayesian Model Averaging of Naive Bayes for Clustering,IEEE Transactions on Systems, Man, and CyberneticsPart B: Cybernetics, VOL. 36, NO. 5, October 2006.

8. Dursun Delen, Glenn Walker, Amit Kadam, Predicting breast cancer survivability:a comparison of three data mining methods, Elsevier publications,Artificial Intelligence in Medicine (2004).

9. A. Kittur, E. Chi, and B. Suh, Crowdsourcing user studies with Mechanical Turk, in Proc. 26th Annu. SIGCHI Conf. Human Factors Comput. Syst., Florence, Italy, Copyright ACM 978-1-60558-011-1/08/04 2008.

10. 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. 18.

11. Thierry Buecheler, Rocky Lonigro, Rudolf M. Fchslin and Rolf Pfeifer, Modeling and Simulating Crowdsourcing as a Complex Biological System: Human Crowds Manifesting Collective Intelligence on the Internet, Artificial Intelligence Laboratory, Department of Informatics, University of Zurich.

12. Karel Dejaeger, Thomas Verbraken, and Bart Baesens, Toward Comprehensible Software Fault Prediction Models Using Bayesian Network Classifiers,IEEE Transactions on Software Engineering, Vol. 39, NO. 2, February 2013

13. Bum Ju Lee, Boncho Ku, Jiho Nam, Duong Duc Pham, and Jong Yeol Kim, Prediction of Fasting Plasma Glucose Status Using Anthropometric Measures for Diagnosing Type 2 Diabetes,IEEE Journal of Biomedical and Health Informatics, Vol. 18, NO. 2, March 2014

14. Luke Gosink, Kevin Bensema, Trenton Pulsipher, Harald Obermaier, Member, IEEE, Michael Henry, Hank Childs, and Kenneth Joy, Member, IEEE, Characterizing and Visualizing Predictive Uncertainty in Numerical Ensembles Through Bayesian Model Averaging,IEEE Transactions on Visualization and Computer Graphics, Vol. 19, NO. 12, December 2013

15. Gaetano Valenti, Maria Lelli, Domenico Cucina, A comparative study of models for the incident duration prediction,Eur. Transp. Res. Rev. (2010) 2:103111 [16] Alfonso Ibez , Concha Bielza,Pedro Larraaga, Cost-sensitive selective naiveBayes classifiers for predicting the increase of the h-index for scientific journals,Neurocomputing135(2014)4252

16. Daniele Soria, Jonathan M. Garibaldi a, Federico Ambrogi, Elia M. Biganzoli, Ian O. Ellis, A non-parametric version of the naive Bayes classifier,Knowledge-Based Systems 24 (2011) 775784

17. 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.

18. Sebastian Raschka, Nave-Bayes and Text Classification Introduction and theory, October, 2004.

19. Jiawei Han and Michelin Kamber, Data Mining Concepts and Technique, second edition.