International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 5, October 2019, pp. 3714~3719
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i5.pp3714-3719

A data mining approach for desire and intention to participate in virtual communities

Özerk Yavuz[1], Adem Karahoca[2], Dilek Karahoca[3]
[1]Department of Computer Engineering, Altinbas University, Turkey
[2]Department of Software Engineering, Engineering and Natural Sciences Faculty, Bahcesehir University, Turkey
[3]Department of Child Development, Health Sciences Faculty, Bahcesehir University, Turkey

Article Info

Article history:

Received Nov 9, 2018
Revised Apr 8, 2019
Accepted Apr 19, 2019

Keywords:

Data mining
Desire and intention to participate in virtual communities
Machine learning
Virtual communities

ABSTRACT

The purpose of this study is to investigate performances of some of the data mining approaches while understanding desire and intention to participate in virtual communities and its antecedents. A research model has been developed following the literature review and the model was tested afterwards. In research part of the study, some of the data mining approaches as JRip, Part, OneR Method, Multilayer Perceptron (Neural Networks), Bayesian Networks have been used. Based on the analysis conducted it has been found out that Multilayer Neural Network had the best correct classification rate and lowest RMSE.

Corresponding Author:

Adem Karahoca,
Department of Software Engineering,
Bahcesehir University,
Faculty of Engineering, Besiktas, Istanbul, 34349, Turkey.
Email: adem.karahoca@eng.bau.edu.tr; akarahoca@gmail.com

1. INTRODUCTION

As virtual community concept emerged during time, new definitions of the term found place in the literature. Porter proposes a virtual community definition that, a virtual community is an aggregation of individuals or business partners who interact around a shared interest, where the interaction is at least partially supported and/or mediated by technology and guided by some protocols or norms [1]. Plant approaches the term from a similar perspective defining a virtual community as a collective group that come together either temporarily or permanently through an electronic medium to enable the interaction of entities, individuals or organizations in a common problem or interest space [2]. In addition to these, Rheingold defines a virtual community as social aggregations that emerge from the Internet when enough people carry on those public discussions long enough, with sufficient human feeling to form webs of personal relationships in cyberspace [3].

The purpose of this study is to investigate performances of some of the data mining approaches while understanding desire and intention to participate in virtual communities and the factors affecting it. For this purpose a model has been developed with the focus on desire and intention to participate to virtual communities and its antecedents. Later following the data gathering phase and pre-processing of the data several data mining approaches have been applied to the data. Some part of the data is used for training purposes whereas remaining is used for testing the model which has been formed following the literature review. Consequently in addition to the studies in the scientific body of knowledge a collaborative and contributive data mining approach is applied to understand desire and intention to participate in virtual communities.

2. RESEARCH METHOD

Data mining can be defined as the process of extracting hidden patterns from large chunks of data. In doing this knowledge discovery, prediction or forecasting can be in the focus of data mining. While knowledge discovery provides us explicit information about the characteristics of the data set predictive modeling provides predictions of future events. As stated by Simoudis, data mining is the process of extracting valid, previously unknown, comprehensible and actionable information from large databases and using it to make business decisions [4]. Data mining borrows approaches from several disciplines as statistics, mathematics or computer science in order to find useful patterns and knowledge from large data sets. As it is indicated in Shearer's crisp-dm model, a data mining process is composed of business understanding, data understanding, data preparation, model building, testing/evaluation and deployment processes. In the following sections some of the data mining approaches used in analyzing the data set will be introduced [5].

2.1. Data gathering and processing

As suggested in literature over 385 observations (425 in our sample in this study) has been found sufficient for the sample size values with an error of 5% and a confidence level of 95% (survey monkey site-sample size calculator). In literature used formula to calculate this has been $n= t^2 \times (p \times q)/e^2$ where n refers to sample size, p refers to proportion, percentage or presence of the study characteristics (in literature it is suggested that when we have no prior values for the proportions to be estimated, we can use p- and q-values as 50%.) q=1-p, e refers to margin of error; t = 1.96 (with 95% confidence level). Based on that, n = 1.962 x 0.5 x 0.5 / 0.052 sample size has been found 384.16 and rounded to 385 [6-7].

Scales used in the study is given in detail. Positive anticipated emotions refer to the pre-factuals hypothesized to influence desires to perform a behavior which can be in the form of positive anticipated emotions or negative anticipated emotions and it's likely to expect its influence on virtual community participation and desire and intention to participate in virtual communities [8]. In the literature, it is pointed out that in general people are in a tendency to expect some return when they share their knowledge. As it is defined by Chiu et al., norm of reciprocity refers to knowledge exchanges that are mutual and perceived by the parties as fair and one of the important factors that leads to knowledge sharing behavior [9].

Perceived usefulness refers to the degree to which a person believes that using a particular system would enhance his or her performance [10, 11]. As it is indicated by Porter, in the technology acceptance model, perceived usefulness and perceived ease of use are the beliefs that are presumed to influence attitudes toward new technology [12]. As it is pointed out in Fishbein and Ajzen's theory of reasoned action, attitudes are formed as a result of the beliefs about the outcomes of performing that act and expected outcomes. If the outcome of performing that behavior seems beneficial to the individual, he/she may participate in that particular behavior [13, 14].

Early definitions of social comparison theory date back to 1954s that started with Festinger's social comparison theory. As stated in the literature according to social comparison theory, there is a drive within individuals to look to outside images in order to evaluate their own opinions and abilities in the sense that it mainly focuses on explaining and understanding tendencies of individuals in evaluating and comparing their own opinions and desires with others which may lead to an self enhancement in individuals' self images. As it is pointed out in literature desires provide the motivation to decide in favor of acting as part of a virtual community. Therefore desire construct has been measured with the help of questions adapted from Dholokia's respective scale [15].

As it is defined by Dholokia, We-Intentions construct used in the model refers to the intentions to participate ingether as a group which is to be a function of both individual (i.e., attitudes, perceived behavioral control, positive, and negative anticipated emotions) and social determinants [15]. Desire and intention to participate in virtual communities refers to the merge of we-intention and desires of Dholakia where desires provide the motivation to decide in favor of acting as part of a virtual community and we intentions stand for the intentions to participate together as a group, to be a function of both individual (i.e., attitudes, perceived behavioral control, positive, and negative anticipated emotions) and social determinants (i.e., subjective norms, group norms, and social identity) [15]. Respective scales have been borrowed empirically from the studies as shown in Table 1.

Table 1. Scales used in the study

Construct	Adapted From
Positive anticipated emotions	Bagozzi, 2002 [8]
Norm of reciprocity	Chiu, 2006 [9]
Perceived Usefulness	Shin, 2008 [10]
Predisposition to Virtual Community Usage	Bagozzi, 2002 [8]
Social Comparison	Chen, 2010 [12]
*Desire And Intention to Participate In Virtual	Dholakia, 2004 [15]
Com. Desires We-Intention	Dholakia, 2004 [15]
	Dholakia, 2004 [15]

*Desire and Intention to Participate in Virtual Communities is the combination of Desires and We Intention Scales

2.2. Data mining methods

As part of the research conducted several data mining approaches have been applied to the data set. Data mining methods can be used more accurately with data preprocessing approaches [16]. Such as normalization of the data, discretization the conutinues data and etc. Brief descriptions of the methods that have been used as follow.

1) *JRip:* JRip implements a propositional rule learner, "Repeated Incremental Pruning to Produce Error Reduction" (RIPPER), as proposed by Cohen, JRip is a rule learner alike in principle to the rule learner Ripper [17]. JRip implements a propositional rule learner, "Repeated Incremental Pruning to Produce Error Reduction" (RIPPER), as proposed by Cohen, JRip is a rule learner alike in principle to the rule learner Ripper [17]. RIPPER rule learning algorithm is an extended version of learning algorithm IREP (Incrementa1 Reduced Error Pruning). It constructs a rule set in which all positive examples are covered, and its algorithm performs efficiently on large, noisy datasets. Before building a rule, the current set of training examples are partitioned into two subsets, a growing set (usual1y 2/3) and a pruning set (usual1y 1/3). The rule is constructed from examples in the growing set. The rule set begins with an empty rule set and rules are added incrementally to the rule set until no negative examples are covered. After growing a rule from the growing set, condition is deleted from the rule in order to improve the performance of the rule set on the pruning examples [18].
2) *PART:* The PART algorithm combines two common data mining strategies; the divide-and-conquer strategy for decision tree learning with the separate-and-conquer strategy for rule learning. The tree building algorithm splits a set of examples recursively into a partial tree. The first step chooses a test and divides the examples into subsets. PART makes this choice in exactly the same way as C4.5. Then the subsets are expanded in order of their average entropy starting with the smallest. The reason for this is that subsequent subsets will most likely not end up being expanded and the subset with low average entropy is more likely to result in a small sub tree and therefore produce a more general rule [19].
3) *OneR:* OneR, generates a one-level decision tree that is expressed in the form of a set of rules that all test one particular attribute. OneR is a method that often comes up with quite good rules for characterizing the structure in data [20]. Pseudo code for 1R is as follow.

   For each attribute,
   For each value of that attribute, make a rule as follows:
   Count how often each class appears
   Find the most frequent class
   Make the rule assign that class to this attribute-value.
   Calculate the error rate of the rules.
   Choose the rules with the smallest error rates [20].
4) *Multilayer Perceptron:* A Multilayer Perceptron is a version of the original perceptron model proposed by Rosenblatt in the 1950s and considered as a type of neural networks (Rosenblatt, 1958). A perceptron (artificial neuron) is a function of several input perceptrons which is formed as a combination of input weights to the hidden layer perceptrons. As stated by Ramchoun in literature multilayer perceptron has one or more hidden layers between its input and output layers, the neurons are organized in layers, the connections are always directed from input layers to output layers and the neurons in the same layer are not interconnected [21]. In this approach hidden layer is a function of the nodes in the previous layer, and the output nodes are a function of the nodes in the hidden layer.
5) *Bayesian Network:* There are no deterministic rules which allow to identify a subscriber as a risk indicator. Graphical models such as Bayesian networks supply a general framework for dealing with uncertainly in a probabilistic setting and thus are well suited to tackle the problem of prediction. Every graph of a Bayesian network codes a class of probability distributions. The nodes of that graph comply

with the variables of the problem domain. Arrows between nodes denote allowed (causal) relations between the variables. These dependencies are quantified by conditional distributions for every node given its parents [22]. A Bayesian network B over a set of variables U is a network structure Bs, which is directed acyclic graph (DAG) over U and set of probability tables $Bp=\{p(u|pa(u))|u \in U\}$ where pa(u) is the set of parents of u in Bs. A Bayesian network represents probability distributions [23, 24].

3. FINDINGS

Reliability of the constructs have been re-assessed and re-evaluated considering suggested lower limit of Cronbach's alpha in literature. As it is shown in Table 2 with the sample size of 425 it has been seen that all Cronbach alpha values for the respective constructs have a value of higher than .70, in other words all the constructs used in the research model are statistically reliable and can be regarded as reliable constructs of the research model [25]. From this reason,

Table 2. Reliability measures of the scales

	Items	Cronbach Alpha
Positive anticipated emotions	7	,918
Norm of reciprocity	2	,798
Perceived Usefulness	3	,886
Predisposition to Virtual Community Usage	4	,857
Social Comparison	3	,869
*Desire and Intention to Participate Virtual Communities	5	,921

In this study, benchmarking of the algorithms of JRip, Part, OneR Method, Multilayer Perceptron, Bayesian Networks have been performed. In testing the research model with each of the data mining approaches 66 percent of the data has been used for the training whereas remaining part of the data set has been used for the testing of the model. Among different data mining approaches JRip had the values (RMSE=0.2913; Precision=N/A; Correct Classification Rate=90.90%; Incorrect Classification Rate=9.09; True Positive Rate=0.909 and False Positive Rate=0.909).

Part had the values (RMSE=0.264; Precision=0.923; Correct Classification Rate=91.60%; Incorrect Classification Rate=8.39; True Positive Rate=0.916 and False Positive Rate=0.839). OneR had the values (RMSE=0.3015; Precision=N/A; Correct Classification Rate=90.90%; Incorrect Classification Rate=9.09; True Positive Rate=0.909 and False Positive Rate=0.909).

Multilayer Perceptron had the values (RMSE=0.2476; Precision=0.921; Correct Classification Rate=93.007%; Incorrect Classification Rate=6.99; True Positive Rate=0.930 and False Positive Rate=0.561) and finally Bayesian Networks had the values (RMSE=0.2873; Precision=0.876; Correct Classification Rate=89.51%; Incorrect Classification Rate=10.49; True Positive Rate=0.895 and False Positive Rate=0.703). Precision values of JRip and OneR method could not been calculated since proportion of instances truly classified of a class divided by the total instances classified in that class have been calculated undefined in the confusion matrix. Among all the algorithms, multilayer perceptron had the most correct classification rate with 93.007 percent, a good true positive rate of 0.930 and a precision 0.921. Part method had a correct classification rate of 91.60 percent, true positive rate of 0.916 and a precision value of 0.923. Multilayer perceptron had the lowest RMSE with a value of 0.24. Comparison of data mining methods used can be seen in Table 3.

Table 3. Comparison of data mining methods used

Method	RMSE	Precision	Correctly Classified %	Incorrectly Classified %	True Positive Rate	False Positive Rate
JRip	0.29	N/A	90.90	9,09	0,90	0,90
Part	0.26	0,92	91.60	8,39	0,91	0,83
OneR Method	0.30	N/A	90.90	9,09	0,90	0,90
Multilayer Perceptron	0.24	0,92	93.00	6,99	0,93	0,56
Bayesian Networks	0.28	0,87	89.51	10,49	0,89	0,70

4. DISCUSSION AND CONCLUSION

In this study, we investigated the factors behind desire and intention to participate in virtual communities following an intensive literature review. This is later followed with the model formation and applying the data mining techniques as suggested in literature. In the analysis part of the study we examined

relationship of positive anticipated emotions, norm of reciprocity, social comparison, predisposition towards virtual community usage and perceived usefulness with desire and intention to participate in virtual communities. In doing so we trained the model using 66 percent of the data of training of the model whereas remaining part for the testing of the model for each approach.

Data mining can be defined as the process of extracting hidden patterns from large chunks of data. In doing this knowledge discovery, prediction or forecasting can be in the focus of data mining. Jrip, part, oner method, Multilayer Perceptron (Neural Networks), and Bayesian Networks have been chosen as the data mining techniques in order to examine desire and intention to participate in virtual communities for this purpose. Among them JRip is a rule learner alike in principle to the rule learner Ripper [17]. The part algorithm combines two common data mining strategies; the divide and conquer strategy for decision tree learning with the separate and conquer strategy for rule learning. Oner generates a one level decision tree that is expressed in the form of a set of rules that all test one particular attribute. A Multilayer Perceptron is a version of the original perceptron model proposed by Rosenblatt in the 1950s and considered as a type of neural networks [26]. A perceptron (artificial neuron) is a function of several input perceptrons which is formed as a combination of input weights to the hidden layer perceptrons which lead them to the output layer. Finaly graphical models such as bayesian networks supply a general framework for dealing with uncertainly in a probabilistic setting and thus are well suited to tackle the problem of prediction.

In this study, we have met our objectives of evaluating and investigating the performances of different data mining techniques for the data set that is being used to understand desire and intention to participate in virtual communities. In addition to the studies in the scientific body of knowledge a collaborative and contributive data mining approach is applied to understand desire and intention to participate in virtual communities. Based on the results, multilayer perceptron had the most correct classification rate with 93.007 percent, a good true positive rate of 0.930 and a precision 0.921. Part method had a correct classification rate of 91.60 percent, true positive rate of 0.916 and a precision value of 0.923. Multilayer perceptron had the lowest RMSE with a value of 0.24. Based on the high correct classification rate and low RMSE measure, multilayer perceptron (neural network) can be considered as an effective method and can be used in understanding desire and intention to participate in virtual communities and its antecedents.

REFERENCES

[1] Porter, C. E., "A typology of virtual communities: A multi-disciplinary foundation for future research," *Journal of Computer-Mediated Communication*, 10(1) Article 3, 2004.

[2] Plant, R., "Online communities," Technology in Society. 26, pp. 51-65, 2004.

[3] Rheingold, H., "The virtual community. USA: MIT Press," 2000.

[4] Simoudis, E. "Reality Check for Data Mining," *IEEE EXPERT*, 11(5), 26-33, 1996.

[5] Shearer, C., "The CRISP-DM model: the new blueprint for data mining," *Journal of Data Warehousing*, vol. 5, pp. 13-22, 2000.

[6] Águila, R.D.M., Ramírez, G.A., "Series: basic statistics for busy clinicians," *Allergol Immunopathol*, 42 (5), pp. 485-492, 2013.

[7] Yau C., "R tutorial with bayesian statistics using openbugs," 2013.

[8] Bagozzi, R.P. & Dholakia U.M., "Intentional social action in virtual communities," *Journal of Interactive Marketing*, 16 (2), pp. 2-21, 2002.

[9] Chiu, C.M., Hsu M., Wang, E., "Understanding knowledge sharing in virtual communities: an integration of social capital and social cognitive theories," *Decision Support Systems*, 42 (3), pp. 1872-1888, 2006.

[10] Shin, D. H., "Understanding purchasing behaviors in a virtual economy: Consumer behavior involving virtual currency in Web 2.0 communities," *Interacting with computers*, 20(4-5), 433-446, 2008.

[11] Davis, F.D., "Perceived usefulness, perceived ease of use and user acceptance of information technology," *MIS Quarterly*, 13 (3), pp. 319–340, 1989.

[12] Porter, E. & Donthu, N., "Using the technology acceptance model to explain how attitudes determine internet usage: the role of perceived access barriers and demographics," *Journal of Business Research*, 59 (9), pp. 999-1007, 2006.

[13] Fishbein, M., Manfredo, M.J., "A theory of behavior change influencing human behavior: theory and applications in recreation, tourism and natural resources management," *Champaign, Illinois: Sagamore Publishing*, 1992.

[14] Ajzen, I. & Fishbein, M., "Understanding attitudes and predicting social behaviour*," Englewood Cliffs, NJ: Prentice Hall*, 1980.

[15] Dholakia, U.M., Bagozzi, R.P. & Pearob, L.K., "A social influence model of consumer participation in network and small group based virtual communities," *International Journal of Research in Marketing*, 21, pp. 241-263, 2004.

[16] Al-Taie, M.Z., Kadry, S., Lucas, J.P., "Online Data Preprocessing: A Case Study Approach," *International Journal of Electrical and Computer Engineering (IJECE)*, 9(4), 2019

[17] Cohen, W. "Fast effective rule induction. In A. Prieditis and S. Russell (eds.)," *Proceedings of the 12th International Conference on Machine Learning*, Lake Tahoe, CA, pp.115–123, 1995.

[18] Sasaki M., Kita K., "Rule based text categorization using hierarchical categories," *IEEE*.
[19] Blackmore, K. & Bossomaier, T.R.J.., "Comparison of See5 and J48.PART algorithms for missing persons profiling," pp. 337-342, 2002.
[20] Frank E. and Witten I.H. "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations," *Morgan Kaufmann Publishers*: San Francisco, CA, 2000.
[21] Ramchoun, H. r., Idrissi, M. m., Ghanou, Y. y., & Ettaouil, M. m. "New Modeling of Multilayer Perceptron Architecture Optimization with Regularization: An Application to Pattern Classification," *IAENG International Journal of Computer Science*, 44(3), 261-269, 2017.
[22] Taniguchi M., Haft M., Hollm´en J., and Tresp V, "Fraud detection in communications networks using neural and probabilistic methods," *In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing,* (ICASSP'98), Volume II, pp. 1241–1244, 1998.
[23] Bouckaert, R. R., "Bayesian network classifiers in Weka," (Working paper series. University of Waikato, Department of Computer Science. No. 14/2004). Hamilton, New Zealand: University of Waikato, 2004.
[24] Burhanuddin, M. A., Ismail, R., Izzaimah, N., Mohammed, A. A-J., Zainol, N., "Analysis of Mobile Service Providers Performance Using Naive Bayes Data Mining Technique," *International Journal of Electrical and Computer Engineering (IJECE)*, 8(6), pp.5153-5161, 2018.
[25] Hair, J., Black, W., Babin, B. & Anderson, R., "Multivariate data analysis. NJ: Prentice Hall," 2010.
[26] Rosenblatt, F., & Cornell Aeronautical Laboratory, "The perceptron: A theory of statistical separability in cognitive systems (Project Para)," Buffalo, N.Y: Cornell Aeronautical Laboratory, 1958.

BIOGRAPHIES OF AUTHORS

Özerk Yavuz holds a PhD in Bussiness Administration and Msc. in Computer Engineering. He is interested in software engineering, computer engineering, data mining, virtual communities, marketing and management. He has published articles in various fields. He has international working experience in several countries and currently working in Altinbas University, Computer Engineering Department in Istanbul.

Adem Karahoca holds a PhD in Software Engineering. He is interested in human–computer interaction, web-based education systems, data mining, big data and management information systems. He has published articles at prestigious journals about data mining applications and business information systems in health, tourism and education.

Dilek Karahoca is a social anthropologist. Holding a PhD in Computer Education and Instructional Technologies, she is interested in human computer interaction, web based education systems, and blended learning methodologies. She has published several articles about use of information systems in health, tourism, and education.