General Bayesian Network in Performing Micro-reality Mining with Mobile Phone Usage Data for Device Personalization

Seong Wook Chae[1], Jungsik Hwang[1], and Kun Chang Lee[2,*]

[1] Sungkyunkwan University, Seoul 110-745, Republic of Korea
[2] SKK Business School and Department of Interaction Science
Sungkyunkwan University, Seoul 110-745, Republic of Korea
{seongwookchae,jungsik.hwang,kunchanglee}@gmail.com

Abstract. Personalization is an emerging issue in the digital age, where users have to deal with many kinds of digital devices and techniques. Moreover, the complexities of digital devices and their functions tend to increase rapidly, requiring careful attention to the questions of how to increase user satisfaction and develop more innovative digital products and services. To this end, we propose a new concept of micro-reality mining in which users' micro behaviors, revealed through their daily usage of digital devices and technologies, are scrutinized before key findings from the mining are embedded into new products and services. This paper proposes micro-reality mining for device personalization and examines the possibility of adopting a GBN (general Bayesian network) as a means of determining users' useful behavior patterns when using cell phones. Through comparative experiments with other mining techniques such as SVM (support vector machine), DT (decision tree), NN (neural network), and other BN (Bayesian network) methods, we found that the GBN has great potential for performing micro-reality mining and revealing significant findings.

Keywords: Reality mining, Personalization, General Bayesian Network, Mobile phone, Usage pattern.

1 Introduction

We are living in a world of digital technologies. Many people use various kinds of digital devices, such as mobile phones and laptop computers. However, since most digital devices tend to become commodities, the success of these devices critically depends on how much manufacturers understand users' micro usage patterns and embed their understanding into the design of the devices' key functions, user interface, outward appearance, and other features. This mining activity is called micro-reality mining and refers to "the system's ability to extract a set of trivial but meaningful rules of action from individuals' behavior data" [1].

[*] Corresponding author.

T.-h. Kim et al. (Eds.): MulGraB/BSBT/IUrC 2012, CCIS 353, pp. 381–388, 2012.

Among the wide array of digital devices in use, we focus on mobile phones because they have become ubiquitous. Therefore, this paper investigates the effect of micro-reality mining for the personalization of mobile phones. By means of micro-reality mining, a cell phone could provide a personalized interface that benefits its user. There are many methods available for micro-reality mining, such as the Bayesian network (BN), the decision tree (DT), the support vector machine (SVM), and the neural network (NN). This paper specifically examines the usefulness of a general Bayesian network (GBN) for performing micro-reality mining on mobile phone usage data. A GBN can provide a set of meaningful causal relationships for the target nodes (or variables). Moreover, GBNs are very useful in suggesting what-if and goal-seeking experiments. This paper assumes that the two key features of GBNs — causal relationships and what-if/goal-seeking experiments — may be critical in adding a sense of reality to the micro-reality mining results. A mobile phone usage data set was selected for the micro-reality mining experiment.

2 Related Works

2.1 Mobile Phone Personalization

Park et al. [2] presented a map-based personalized recommendation system based on user preference. A BN was used to model user preferences using profile and context information obtained from mobile devices. Yan and Chen [3] introduced the AppJoy system, which provides personalized application recommendations to its users. Lee and McKay [4] investigated a personalized keypad layout for mobile phones. They employed a genetic algorithm (GA) to match a suitable keypad layout to each user. GA-based optimization of mobile keypads can also be found in the work of How and Kan [5] and Moradi and Nickabadi [6]. Olwal et al. [7] investigated customization of mobile devices for elderly users and presented a prototype framework, OldGen. Mishra et al. [8] presented a method for personalized search on mobile phones based on the user's location. Rosenthal et al. [9] introduced personalized mobile phone interruption models that automatically adjust phone volume to avoid phone interruptions.

2.2 Micro-reality Mining

The term "micro-reality mining" was first suggested by Chae et al. [1], who described it as a system's capabilities to extract trivial but meaningful information from data. The concept is based on the fact that such trivial rules of action determine an individual's macro behavior. According to Fayyad et al. [10], data mining is the process of applying specific algorithms to find patterns or construct models from data and plays a critical role in the process of knowledge discovery [10]. There are many representations of data mining, including decision trees, artificial neural networks, and rule bases. There are also a variety of techniques for data mining, such as classification, regression, density estimation, and clustering [11].

2.3 Bayesian Network

A Bayesian network is a strong tool for modeling decision making under conditions of uncertainty [12-13], meaning that decision makers facing uncertainty can rely on a BN to get robust decision support with respect to the target problem.

A BN is composed of nodes, links, and conditional probability tables and applies the Bayes rule to probability theory. This rule specifies how existing beliefs should be modified mathematically with the input of new evidence. A BN consists of a directed acyclic graph (DAG) and a corresponding set of conditionals (conditional probability tables) [14]. The DAG shows the causal relationship between the variables and the nodes, where each node represents a variable that has an associated conditional probability distribution.

Researchers have studied BNs in earnest since the naïve Bayesian network (NBN) was presented. A NBN has the simplest shape of all BNs —a class node linked with all the other nodes — and does not explain the causal relationships between the child nodes. It has been shown to have relatively high accuracy [15]. NBNs have been shown to be surprisingly accurate in many domains when compared with alternatives such as rule learning, decision tree learning, and instance-based learning [16]. The tree-augmented NBN (TAN) is an extended version of the NBN in which the nodes form the shape of a tree. Experiments have indicated that it significantly outperforms the NB in classification accuracy [17]. To improve classification performance, Cheng and Greiner [18] suggested a general Bayesian network (GBN). In comparison with NBNs and TANs, GBNs can be structured very flexibly. Given a class node (or target node), a set of relevant explanatory nodes can be linked with each other [19].

2.4 Other Classifiers

There are many other classifiers used in data mining, including NN, SVM, and DT classifiers. A NN classifier mimics a neuronal structure of the human brain that performs classification with knowledge stored in connections between nodes in the network [20]. The SVM classifier is "a binary classifier which looks for an optimal hyperplane as a decision function in a high-dimensional space" [21]. A DT classifier is also a binary tree in which every non-leaf node is related to a predicate [22].

3 Research Methodology and Experiment

3.1 Data

In our study, we used a data set collected by the Reality Mining Group at MIT Media Lab [23]. From the many variables provided, we carefully selected those related to mobile phone usage patterns. Then, we organized the variables into individual units for one month's worth of data instead of using the full nine months of data. Consequently, the total number of data elements in our analysis was 80. The 20 variables included in the analysis are summarized in Table 1.

Table 1. Variables

Variable Name	Description
*call_N_in	Number of incoming calls
*call_N_missed	Number of missed calls
*call_N_out	Number of outgoing calls
*call_N_inOut	Sum of number of incoming and outgoing calls
*call_N_total	Sum of number of incoming, missed, and outgoing calls
*call_du_in	Sum of duration (seconds) of incoming calls
*call_du_out	Sum of duration (seconds) of outgoing calls
*call_du_inOut	Sum of duration of incoming and outgoing calls
**call_du_avg_in	Average duration of an incoming call
**call_du_avg_out	Average duration of an outgoing call
**call_du_avg_inOut	Average duration of a call
**vcDuOutPref	call_du_out / (call_du_out+ call_du_in) * 100
**vcOutter	call_N_out / call_N_in
**vcIncommer	call_N_in / call_N_out
**vcOutPref	call_N_out / (call_N_out+ call_N_in) * 100
*smIncoming	Number of incoming SMS
*smOutgoing	Number of outgoing SMS
*smTotal	Number of incoming and outgoing SMS
**smOutPref	smOutgoing / (smOutgoing + smIncoming) * 100
*****caVCpref**	call_N_out / (call_N_out+ smOutgoing) * 100

* Observed variable.

** Manipulated variable.

*** Target variable.

3.2 Experiments

We used WEKA version 3.6.7 (Waikato Environment for Knowledge Analysis) [24] to extract rules from the dataset. Seven different classifiers were examined: NBN, TAN, GBN-K2, GBN-Hill Climb (GBN-HC), NN, DT, and SVM. To construct the GBN-K2 and GBN-HC classifiers, we set the maximum number of parent nodes to 5. In addition, the BAYES scoring metric was used for Bayesian classifiers. To construct other classifiers, the default settings in WEKA were used. Because WEKA can handle nominal variables only, we discretized a range of numeric variables by using 3 equal-width bins. Table 2 shows the prediction accuracies of the classifiers.

Table 2. Prediction accuracies of classifiers

	NBN	*TAN*	*GBN-K2*	*GBN-HC*	*NN*	*SVM*	*DT*
Mean(S.D.)	85.30 (5.96)	88.42 (4.89)	**89.25** (4.74)	86.87 (6.95)	86.48 (4.81)	86.77 (5.18)	**84.20** (6.32)

To determine classification performance, a one-way ANOVA followed by a post-hoc test using Dunnet T3[1] procedures was conducted on the mean scores of the 7 classifiers. Table 2 shows the mean values of the prediction accuracy[2] for the 7

[1] Tests of homogeneity of variances (Levene's test) showed that in all cases variances were significantly heterogeneous (Levene statistic = 5.134, $p < 0.001$).

[2] The classification performance was measured 100 times.

classifiers. ANOVA yielded statistically significant mean differences (F(6,693) = 9.376, $p < 0.001$), indicating that the mean accuracy performance was significantly different among the 7 classifiers. Based on the numbers in Table 2, the GBN-K2 classifier was regarded as demonstrating the best classification performance (m = 89.25, S.D. = 4.74), and the TAN classifier was ranked second (m = 88.42, S.D. = 4.89); the DT classifier demonstrated the worst performance (m = 84.20, S.D. = 6.32). However, the post-hoc test results indicated that there was no significant difference among the GBN-K2, TAN, and GBN-HC classifiers. In sum, the results from ANOVA demonstrated that the GBN-K2 and TAN classifiers demonstrated significantly better performance than the other classifiers.

4 Simulation

4.1 Selecting Classifier for Simulation

It is necessary to select an appropriate classifier to simulate the model. In the previously described classifier accuracy test, both the GBN-K2 and TAN classifiers demonstrated better performance than any other classifiers. For research purposes, we selected the GBN-K2 classifier for our simulation because a GBN is a full-fledged BN in which causal relationships between the class node and all other nodes are flexibly formulated. These characteristics of a GBN make it possible to conduct the scenario simulation in various ways while making interpretation of the results more fruitful.

Fig. 1 illustrates the constructed Bayesian network. The causal relationships between variables are depicted as directed arches and the probability distribution of each node is shown. As shown in Fig. 1, there were causal relationships between the target node caVCpref and descriptive variables such as call_N_in, call_N_missed, call_N_out, call_N_inOut, call_N_total, call_du_in, call_du_out, call_du_inOut, call_du_avg_in, call_du_avg_out, call_du_avg_inOut, vcDuOutPref, vcOutter, vcIncommer, vcOutPref, smIncoming, smOutgoing, smTotal, and smOutPref. In addition, the causal relationship between the descriptive variables was observable, demonstrating the advantages of using a general Bayesian network.

4.2 Sensitivity Analysis

To examine the possibility of employing micro-reality mining for mobile phone personalization, we simulated the models constructed from the selected classifier, GBN-K2. We then considered two scenarios, what-if and goal-seeking, and performed a sensitivity analysis for each, taking advantage of the causal relationships suggested by Fig. 1.

Scenario 1. (What-if analysis) *If call_N_in is set to low, smIncoming is set to high, and no other variables are changed, what changes occur in caVCpref and the other variables?*

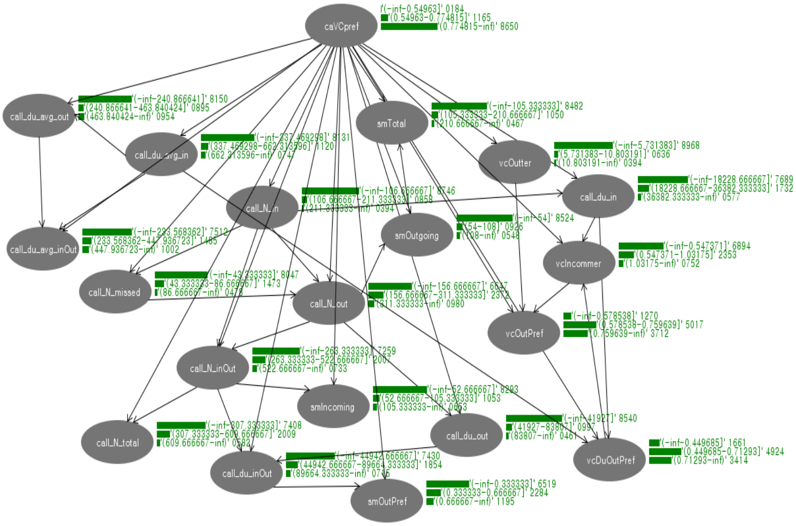

Fig. 1. Constructed Bayesian Network Model with caVCpref as a target node

In Scenario 1, we were able to obtain an idea of an individual's outgoing call preference based on the individual having a small number of incoming calls and a large number of incoming SMS (text) messages. Before we changed the values of call_N_in and smIncoming, the target node (degree of outgoing call preference, or caVCpref) was high, as shown in Fig. 1, indicating that, in general, people are more likely to make an outgoing call than to send a text message. These results indicate that people who have a small number of incoming calls and a large number of incoming SMS messages tend to make voice calls rather than sending texts.

Scenario 2. (Goal-seeking analysis) *To make caVCpref high, what other factors should be changed?*

We also conducted a goal-seeking analysis using the following scenario: For a user who prefers sending SMS messages to calling (low degree of outgoing call preference), what mobile usage patterns are observed? When the target node (caVCpref) was manipulated according to the scenario, the probability distribution of other variables changed. We focused primarily on two variables: sum of duration of incoming calls (call_du_in) and sum of duration of outgoing calls (call_du_out). A user who prefers sending SMS messages to calling might use a mobile phone more actively and frequently compared to other users. This type of user would have longer duration incoming and outgoing calls compared with users who prefer calling to sending SMS messages.

The usage pattern, which seemed meaningless before analysis, demonstrated that there was indeed a pattern or a model. Micro-reality mining can be used to extract several rules. For example, Soto et al. [25] investigated the relationship between individuals' socioeconomic levels and their cell phone records, constructing predictive models using SVM and random forests. As Soto et al. [25] stated, trivial acts can indeed determine macro user behavior.

5 Concluding Remarks

This paper investigated the effect and applicability of micro-reality mining for device personalization and the possibility of employing a GBN as a means of deriving causal relationships from mobile phone usage patterns. The results demonstrated that the valid causal relationships of the GBN can be used to provide new insights into users' micro behaviors when using mobile phones. In addition, these users' micro behaviors could be analyzed to predict users' macro behaviors.

Consequently, the present study not only showed the importance of micro-reality mining, but also suggested the implications for future study. By means of micro-reality mining, individuals' trivial behavior data could be used to reveal meaningful rules of actions. These insights, in turn, can be used in the personalization of mobile devices. By analyzing user's cell phone usage patterns (micro behaviors), the mobile devices can be tailored to meet user's requirements (macro behaviors). Especially, we hope that potentials of GBN may be recognized much more in future studies of personalizing smartphone user-interface, and upgrading user satisfaction.

Acknowledgment. This research was supported by WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. R31-2008-000-10062-0).

References

1. Chae, S.W., Hahn, M.H., Lee, K.C.: Micro Reality Mining of a Cell Phone Usage Behavior: A General Bayesian Network Approach. In: 2011 International Conference on Ubiquitous Computing and Multimedia Applications (UCMA), pp. 119–122 (2011)
2. Park, M.-H., Hong, J.-H., Cho, S.-B.: Location-Based Recommendation System Using Bayesian User's Preference Model in Mobile Devices. In: Indulska, J., Ma, J., Yang, L.T., Ungerer, T., Cao, J. (eds.) UIC 2007. LNCS, vol. 4611, pp. 1130–1139. Springer, Heidelberg (2007)
3. Yan, B., Chen, G.: AppJoy: personalized mobile application discovery. In: Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services (MobiSys 2011), pp. 113–126. ACM, New York (2011)
4. Lee, J., McKay, B.: Optimizing a Personalized Cellphone Keypad. In: Lee, G., Howard, D., Ślęzak, D. (eds.) ICHIT 2011. LNCS, vol. 6935, pp. 237–244. Springer, Heidelberg (2011)
5. How, Y., Kan, M.-Y.: Optimizing Predictive Text Entry for Short Message Service on Mobile Phones. In: Human Computer Interfaces International, HCII 2005 (2005)
6. Moradi, S., Nickabadi, A.: Optimization of Mobile Phone Keypad Layout Via Genetic Algorithm. In: Information and Communication Technologies, ICTTA 2006, 2nd edn., pp. 1676–1681 (2006)

7. Olwal, A., Lachanas, D., Zacharouli, E.: Oldgen: Mobile Phone Personalization for Older Adults. In: Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems (CHI 2011), pp. 3393–3396. ACM, New York (2011)
8. Mishra, V., Arya, P., Dixit, M.: Improving Mobile Search through Location Based Context and Personalization. In: 2012 International Conference on Communication Systems and Network Technologies, pp. 392–396 (2012)
9. Rosenthal, S., Dey, A.K., Veloso, M.: Using Decision-Theoretic Experience Sampling to Build Personalized Mobile Phone Interruption Models. In: Lyons, K., Hightower, J., Huang, E.M. (eds.) Pervasive 2011. LNCS, vol. 6696, pp. 170–187. Springer, Heidelberg (2011)
10. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The Kdd Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM 39(11), 27–34 (1996)
11. Heckerman, D.: Bayesian Networks for Data Mining. Data Mining and Knowledge Discovery 1(1), 79–119 (1997)
12. Cowell, R.G., Dawid, A.P., Lauritzen, S.L., Spiegelhalter, D.J.: Probabilistic Networks and Expert Systems. Springer, New York (1999)
13. Jensen, F.V.: Bayesian Networks and Decision Graphs. Springer, New York (2001)
14. Shafer, G.: Probabilistic Expert Systems. Society for Industrial and Applied Mathematics, Philadelphia (1996)
15. Langley, C.J., Holcomb, M.C.: Creating Logistics Customer Value. Journal of Business Logistics 13(2) (1992)
16. Langley, P., Sage, S.: Induction of Selective Bayesian Classifiers. In: Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence, pp. 339–406 (2006)
17. Friedman, N., Geiger, M., Goldszmidt, M.: Bayesian Network Classifiers. Machine Learning 29(2), 131–163 (1997)
18. Cheng, J., Greiner, R.: Learning Bayesian Belief Network Classifiers: Algorithms and System. In: 14th Canadian Conference on Artificial Intelligence, pp. 141–151 (2001)
19. Silander, T., Myllymäki, P.: A Simple Approach for Finding the Globally Optimal Bayesian Network Structure. In: Proceedings of 22nd Conference on Uncertainty in Artificial Intelligence (2006)
20. Liao, K., Paulsen, M.R., Reid, J.F., Ni, B.C., Bonifacio-Maghirang, E.P.: Corn Kernel Breakage Classification by Machine Vision Using a Neural Network Classifier. Transactions of the ASAE (American Society of Agricultural Engineers) 36(6), 1949–1953 (1993)
21. Cristianini, N., Shawe-Taylor, J.: Introduction to Support Vector Machines. Cambridge University Press (2000)
22. Jin, R., Agrawal, G.: Effcient Decision Tree Construction on Streaming Data. In: SIGKDD Conference on Knowledge Discovery and Data Mining (KD), pp. 571–576 (2003)
23. Eagle, N., Pentland, A.S.: Reality Mining: Sensing Complex Social Systems. Personal and Ubiquitous Computing 10(4), 255–268 (2006)
24. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The Weka Data Mining Software: An Update. ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) Explorations Newsletter 11(1), 10–18 (2009)
25. Soto, V., Frias-Martinez, V., Virseda, J., Frias-Martinez, E.: Prediction of Socioeconomic Levels Using Cell Phone Records. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) UMAP 2011. LNCS, vol. 6787, pp. 377–388. Springer, Heidelberg (2011)