Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 100 (2016) 782 – 789

Procedia
Computer Science

Conference on ENTERprise Information Systems / International Conference on Project MANagement / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2016, October 5-7, 2016

Utilizing longitudinal data to build decision trees for profile building and predicting eating behavior

Gerasimos Spanakis[a,*], Gerhard Weiss[a], Bastiaan Boh[b], Vincent Kerkhofs[b], Anne Roefs[b]

[a]*Department of Data Science and Knowledge Engineering, Maastricht University, 6200MD, The Netherlands*
[b]*Faculty of Psychology and Neuroscience, Maastricht University, 6200MD, The Netherlands*

Abstract

In this paper a framework for warning people when they are at risk of unhealthy eating is presented. Data is collected trough a mobile application called "ThinkSlim" which was developed for the purpose of studying eating behavior using Ecological Momentary Assessment (EMA) principles. Data is converted in order to allow early prediction of healthy and unhealthy eating events and a decision tree algorithm taking into account the longitudinal structure of the dataset is utilized to predict healthy versus unhealthy eating events. Rules that are derived from this decision tree are used to cluster users to groups based on the rule triggering frequencies. Groups created are used for providing users with semi-tailored feedback and are analyzed providing useful insights regarding the conditions that lead to unhealthy eating among different participants allowing for building different eating profiles.

Peer-review under responsibility of the organizing committee of CENTERIS 2016

Keywords: ecological momentary assessment, decision trees, user profiling

1. Background & Introduction

Nowadays obesity is considered to be a pandemic due to its prevalence around the world[1] and treatments are generally not successful. They lead to weight loss in the short term, but weight is often regained in the longer term.[2] With the rise of mobile technology and the internet, it has become possible to provide treatment frameworks like Ecological Momentary Intervention (EMI) which uses a combination of real-time assessment (Ecological Momentary Assessment, EMA) and treatment. Therefore, EMI allows the provision of (indefinite) treatment in the natural environment.[3] To accomplish this, assessment and treatment is conducted and provided via a mobile platform, such as a smartphone. The advantage over traditional treatment is that EMI does not necessarily involve therapist contact but observations made in daily life are used as input to guide therapy-based techniques and progress. Therefore, EMI is most suitable in combination with a well-defined and structured intervention protocol.

[*] Corresponding author: Tel. +31(0)43 38 83916
Email address: jerry.spanakis@maastrichtuniversity.nl (Gerasimos Spanakis)

Peer-review under responsibility of the organizing committee of CENTERIS 2016
doi:10.1016/j.procs.2016.09.225

"ThinkSlim" is an iPhone application developed to collect real-life data from people and help them detect their unhealthy eating events before they occur. The application makes use of EMA concepts which provides us lots of data with a rich longitudinal hierarchical structure. More specifically, through an elaborate questionnaire system, participants provide information in-situ and subsequently appropriate feedback is provided to the participants when a relevant event occurs. All collected data is stored locally and synchronized with a dedicated server for further analysis.

EMA research methods use mobile technology (diaries, PDAs, smartphones et.c.) to collect repeated measurements on the same unit (i.e. humans, plants, samples depending on the study) over time, e.g. experiencing craving is measured again and again on the same subject. Classical statistics often assume that observations are drawn from the same general population and are independent and identically distributed.[4] This assumption is not applicable to EMA data and most machine learning algorithms do not take this into account when treating these data.[5] There have been some efforts to apply decision tree based methods to EMA data[6] to overcome dependencies between data but with limited applications. Other approaches tried to introduce a random factor, but they are only applied to regression trees.[7,8,9]

So far, limited studies have been conducted utilizing EMI for obesity.[10] This study has shown that at the end of the intervention, participants intake of healthful food increased, and that the intervention was considered acceptable by participants. More research is necessary to improve insights into the efficacy of EMI for obesity.

In this paper, we present a framework which takes advantage of the longitudinal structure of the data and predicts under which conditions participants are more likely consume unhealthy food. The proposed methodology utilizes decision trees to derive rules that represent in a very simple way the probability of eating behavior of participants. Decision Trees were chosen as a model (instead of other classification algorithms) since they are simple to understand and interpret: The reasoning for every decision is easily explained and the derivation of the rules provides the background for assisting with case-tailored feedback. More specifically, extracted rules' occurrences are used as vectors representative of the eating behavior. Subsequently, participants will be clustered together according to similarity of the vectors of rules that predict eating behavior. Each cluster of participants is thus represented by a ruleset that is used during the EMI to provide therapeutic feedback when at risk for unhealthy eating. Each group is represented by a ruleset which is used to provide feedback to participants in risky moments. The rest of the paper is organized as follows: Section 2 presents the data collection and preparation process. The proposed algorithm is described in Section 3, followed by experimental data in Section 4. Finally, Section 5 concludes the paper.

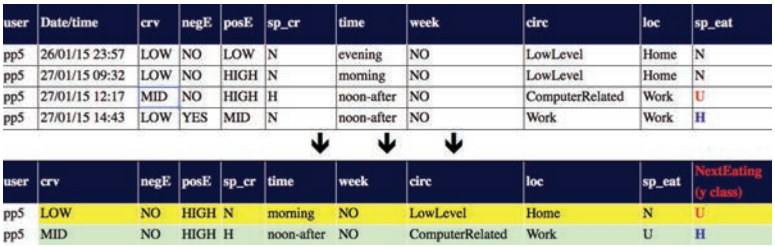
2. Data Collection and Transformation

In the "ThinkSlim" application EMA is performed in two ways: (a) Random sampling: Limited input is requested at pseudo-random time points throughout the day (pseudo-random means that the waking day is divided into on average 8 2-hour timeframe boxes, and assessments occur at random times within each box). Every day subjects are randomly notified by a beeper (random sampling) between 0730 and 2230 (approximate times since participant's actual bedtime habits are taken into account) with an interval of two hours and (b) Event sampling: participants are instructed to use the application immediately prior to eating something, filling a similar questionnaire to random sampling moments with additional information regarding the food items that were about to be consumed. This process results in an average of 10 responses (including random samples and eating events) per participant per day. The dataset is multi-level and complex containing information about users and their eating events, emotions, circumstances, locations, thoughts, food desires (cravings) for several time moments. More information about the study can be found in.[11] Based on exploratory analysis statistics, data (numeric & free text) is discretized and the possible values for each attribute are shown in Table 1 (along with any other categorical attributes). It should be noticed that healthy versus unhealthy eating is based on the choices of food products made by the participants, so it does not refer to irregular eating, disorders, etc, since the purpose of the study is to monitor eating behavior in general.

After this process, data is organized into users and timestamps which contain the information available in Table 1. Each data point is used to predict whether the next data point (provided that they both occur on the same day and obviously, derive from the same user) will be a healthy or an unhealthy eating event. Figure 1 shows an example of how data points (belonging to user "pp5") are converted and combined in a time-lagged fashion so as to enable early prediction using a classification algorithm.

Attribute	Short	Cardinality	Discretized values	Details
Craving/Food Desire	crv	3	Low, Mid, High	
Negative Emotions	negE	2	No, Yes	sad, bored, stressed, angry
Positive Emotions	posE	3	Low, Mid, High	happy, relaxed
Location	loc	6	Home, School, Traveling, Work, Social, Other	
Circumstances (Activities)	circ	10	ComputerRelated	Phone / Internet / Computer
			Eating	Eating / Non-social drinking
			HighLevelIn	Preparing food, cleaning, sanitary, etc.
			HighLevelOut	Exercising, hobby, leisure, shopping, etc.
			LowLevel	Relaxing, waiting, lying in bed, etc.
			WatchingTV	
			Reading	Studying, thinking, etc.
			Socializing	Having a drink, etc.
			Outdoors	traveling, etc.
			Working	administration, work activities, etc.
Time of day	time	3	morning, noon-afternoon, evening	
Weekend	week	2	NO, YES	
Specific Craving	sp_cr	3	N, H, U	Nothing, Healthy, Unhealthy
Specific Eating	sp_eat	3	N, H, U	Nothing, Healthy, Unhealthy

Table 1: ThinkSlim dataset attributes



user	Date/time	crv	negE	posE	sp_cr	time	week	circ	loc	sp_eat
pp5	26/01/15 23:57	LOW	NO	LOW	N	evening	NO	LowLevel	Home	N
pp5	27/01/15 09:32	LOW	NO	HIGH	N	morning	NO	LowLevel	Home	N
pp5	27/01/15 12:17	MID	NO	HIGH	H	noon-after	NO	ComputerRelated	Work	U
pp5	27/01/15 14:43	LOW	YES	MID	N	noon-after	NO	Work	Work	H

user	crv	negE	posE	sp_cr	time	week	circ	loc	sp_eat	NextEating (y class)
pp5	LOW	NO	HIGH	N	morning	NO	LowLevel	Home	N	U
pp5	MID	NO	HIGH	H	noon-after	NO	ComputerRelated	Work	U	H

Fig. 1: Data conversion example for early prediction

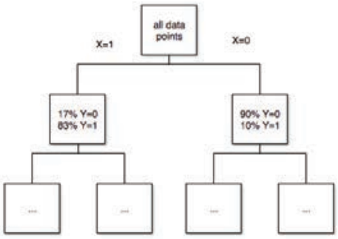
3. Proposed Framework

In this Section the proposed framework is presented: Firstly, the decision tree construction and the derivation of the rules is introduced, secondly, the utilization of rules in the individual user profile construction is presented and finally, the adaptive feedback module that provides users with warnings over possible unhealthy eating moments is described.

3.1. Decision Tree building and Rule induction

Using the data points of Figure 1 as observations, we want to predict under which conditions (i.e. combinations of attributes) participants are led to unhealthy eating. In order to (recursively) build a decision tree, we need to select the “most important” attribute to “split” the data.[12] In our case we select Information Gain (IG) but the branching is performed in a way that takes into account the longitudinal structure of the data.

First, the attribute with the largest Information Gain(IG) is selected. Then, if C is the dominant class (for each new node) we define $Z_k = +$ for every user k if the number of observations (in that node) with $Y = C$ is greater or equal than the number with $Y \neq C$. Otherwise, $Z_k = -$. We form a contingency table with the 2^k patterns of Z as columns and the attribute splits as rows and compute the significance using an independence test (Fisher Test). If the test is positive, then the associated variable is selected for splitting and we continue building the tree. If not, the variable with the second best IG is selected and the process is repeated. An example of this process can be found in Figure 5. In this Figure, we assume a small dataset of 17 data samples and we want to assess whether attribute X is suitable for branching. Firstly, we construct a contingency table for computing the IG. This Table is the 2×2 table on top of Figure 2b. IG based on these numbers is 0.2931. Then, we form the contingency table based on the previous process which leads to the bottom table of Figure 2b. The significance of this table is computed using Fisher Test and the result of the test is positive (p-value 0.0009791), so attribute X will be selected for branching. In Figure 2a the branching can be seen and how it improves the splitting of data points in regard to the outcome Y. Provided we are looking for more accuracy we can repeat the same process recursively for the two new created nodes, which usually is the case for large datasets.



(a) Decision Tree node creation

	Y=0	Y=1	total
X=0	10	1	11
X=1	1	5	6
total	11	6	17

user 1	-	-	-	-	+	+	+	+
user 2	-	-	+	+	-	-	+	+
user 3	-	+	-	+	-	+	-	+
X=0	0	0	0	0	0	8	0	10
X=1	0	0	0	0	4	5	0	0

(b) Top: Contingency table to compute Information Gain, Bottom: Contingency table to assess the node creation from the user perspective

Fig. 2: Explanatory process of building the decision tree

3.2. *User profiling*

We apply the algorithm described in the previous Section to extract (suppose N) significant rules that indicate what combinations of states of variables (e.g. scoring high on craving + being at home + feeling bored + feeling calm) are predictive of unhealthy or healthy eating. Both healthy and unhealthy eating are considered in order to demonstrate the conditions that lead to unhealthy eating compared to healthier options and also for better assessment of eating behavior.

In order to be able to construct profiles of eating behavior based on the rules, the data samples of all participants (suppose P) are checked to compute the rule triggering frequency. More specifically, each participant is represented by a N-dimensional vector (rule vector), where each component represents a rule. The value of the component represents the frequency of occurrence of that rule for the participant.

Next, participants are compared based on their rule vectors (by taking their Euclidean distance) and are grouped together using a standard Hierarchical Agglomerative Clustering (HAC) algorithm.[13] This results in M groups of participants (M is determined by standard evaluation of the clustering results), and each group is described by a rule vector (similar to the participant vectors) that is representative of the rule frequencies within the group. Finally, each group is represented by a ruleset that describes 80% of the eating behaviour of participants in the group (thus removing rules with low occurrence and keeping only those with high predictive value). This process is described in Figure 3.

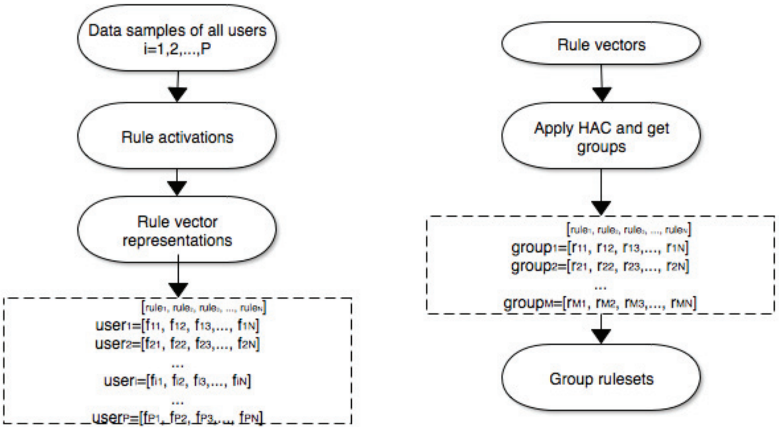
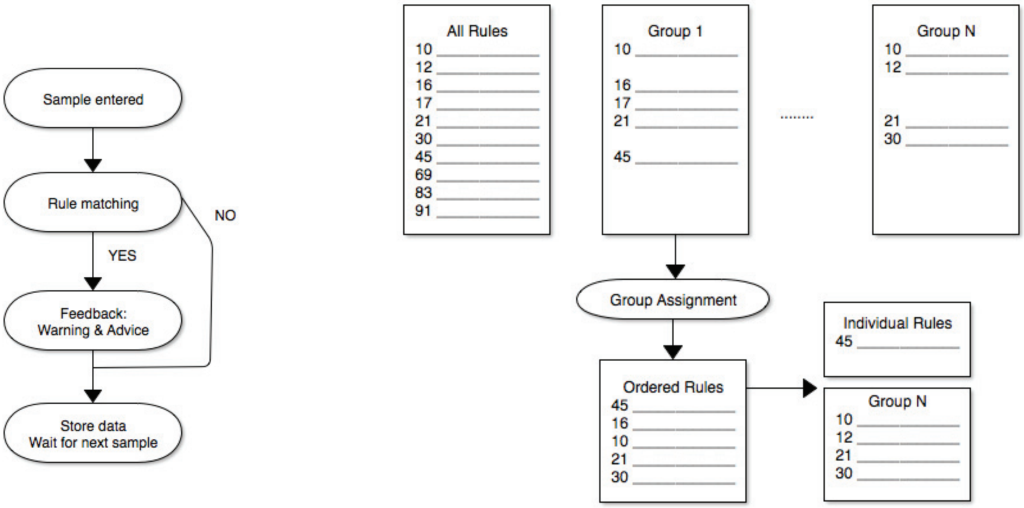

Data samples of all users i=1,2,...,P
Rule activations
Rule vector representations
[rule1, rule2, rule3, ..., ruleN]
user1=[f11, f12, f13,..., f1N]
user2=[f21, f22, f23,..., f2N]
...
useri=[fi1, fi2, fi3,..., fiN]
...
userP=[fP1, fP2, fP3,..., fPN]
Rule vectors
Apply HAC and get groups
[rule1, rule2, rule3, ..., ruleN]
group1=[r11, r12, r13,..., r1N]
group2=[r21, r22, r23,..., r2N]
...
groupM=[rM1, rM2, rM3,..., rMN]
Group rulesets

Fig. 3: Group ruleset construction

3.3. *Towards tailored feedback*

The adaptive feedback module of the application provides participants with feedback if they are at risk for overeating. Detection of these risky moments is based on the answers provided by the participant on the EMA items. Every new (random) sample that is completed by a participant, is checked for a match with one of the pre-existing rules (using the decision tree) and provided there is a match, the participant receives a warning and a behavorial advice via the application. Note that these feedback messages can only occur after a random sample is completed by the participant, and will only occur when the application detects that the participant is likely to eat something that is considered unhealthy in the time period directly following the random sample. This process is shown in Figure 4a.

To provide a degree of tailoring for feedback for new participants, their samples are analyzed over a certain time frame (e.g. one week) and a user profile is obtained by matching the user to the predefined set of M groups obtained with the process of the previous Section. Each group has its own set of rules, where a rule is a combination of variables that has statistically been shown to lead to unhealthy eating for users with the common group profile. Comparison between group vectors and user vectors is possible through the same distance measure (Euclidean distance) used for comparing participants and reveals which rule set will be assigned to which participant (obviously we select the group which has the smallest Euclidean distance to the user rule vector). To allow for individual tailoring, rules that have shown to be statistically important to the participant, but do not belong to the group rule set, are included in the active set of rules. This process is shown in Figure 4b. More information about the study protocol can be found in.[14]



(a) Rule activation and feedback

(b) Tailored feedback to the user

Fig. 4: Adaptive feedback process

4. Experimental Results

Given a sample of $N = 57$ obese participants, we extracted 65 significant rules (36 leading to healthy eating and 29 to unhealthy) using the algorithm described in Section 3.1. An example of what a decision tree looks like can be seen in Figure 5. Note that the real tree structure is far more complex and dense. Given the decision tree structure, we follow every path that leads from root to a leaf and infer one rule per leaf. On each node the split condition can be seen: If it is "true" (i.e. "yes") we take the left branch, otherwise we take the right branch. For example, in the sample tree of Figure 5, the rule corresponding to the far-right leaf is induced by following the red line:

IF CIRCUMSTANCES = {ComputerRelated,Outdoors,Reading,Socializing,Watching TV}
AND SPECIFIC_CRAVING={U}
→ NEXT_EATING={U}

This rule is activated when user completes a sample and e.g. has craving for something unhealthy and is watching TV, resulting in a warning about a "possible" unhealthy eating event.

More examples on the extracted rules can be found in Figure 6.

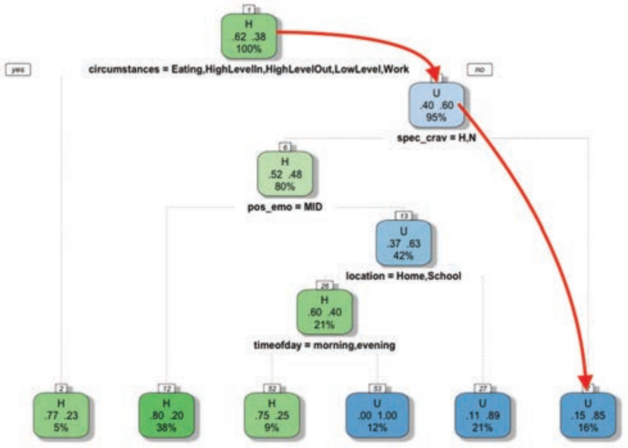


Fig. 5: Decision Tree Example

IF SPECIFIC_CRAVING={H,N}
AND SPECIFIC_EATING={H,U}
AND CIRCUMSTANCES={Eating,HighLevelIn,LowLevel,Socializing,Watching TV}
AND TIMEOFDAY={noon-afternoon, evening}
→ NEXT_EATING={U}

IF SPECIFIC_CRAVING={U}
AND TIMEOFDAY={morning}
AND SPECIFIC_EATING={N}
AND LOCATION={Outdoors,School,Social}
→ NEXT_EATING={U}

IF SPECIFIC_CRAVING={H,N}
AND SPECIFIC_EATING={N}
AND TIMEOFDAY={evening}
AND CRAVING={Low,Mid}
AND CIRCUMSTANCES={ComputerRelated,Reading,Watching TV,Work}
AND LOCATION={Home,Other,Work}
AND POSITIVE_EMOTIONS={Low}
AND NEGATIVE_EMOTIONS={Yes}
→ NEXT_EATING={U}

IF SPECIFIC_CRAVING={H,N}
AND SPECIFIC_EATING={N}
AND TIMEOFDAY={morning, noon-afternoon}
AND LOCATION={Outdoors, Traveling, Other}
AND POSITIVE_EMOTIONS={Mid}
AND WEEKEND={Yes}
→ NEXT_EATING={U}

Fig. 6: Rule examples

After the extraction of the rules the profiling process is taking place. Every participant is represented by a 65-dimensional-vector and using a Hierarchical Agglomerative Clustering (HAC) users are clustered. The results of HAC can be found in Figure 7. Since clustering is an unsupervised algorithm the optimal number of groups has to be decided using standard criteria.[15] In our case, multiple criteria suggested that the optimal number of clusters is 6 (groups are denoted with different color in Figure 7).

In Table 2 some characteristics for the groups created can be found. From this Table, it becomes apparent that group 2 features the most healthy-eating participants, since they tend to activate less unhealthy rules than any other group (5.30%) and this is the reason of the low rate of triggers per day (0.42). This is also supported by the fact that the percentage of unhealthy rules that are triggered is much lower than the percentage of healthy rules (19%). In contrast to these finding, group 6 features the participants which activated mostly unhealthy rules (52.4%) and they

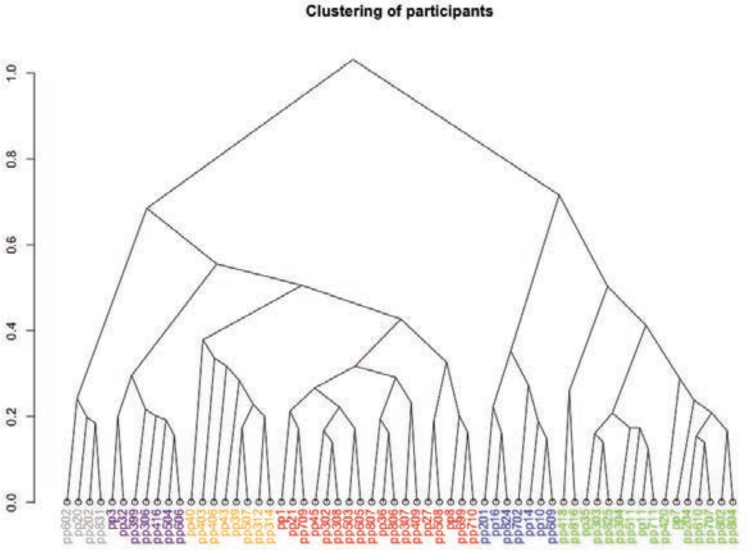


Fig. 7: Clustering process
Grey: Group 4, Purple: Group 5, Yellow: Group 6, Red: Group 1, Blue: Group 2, Green: Group 3

also trigger almost 2 warnings per day (on average).

Group	# users	# active rules	% of U rules	average % of rule triggering	average # of warnings per day
1	18	14	42%	11.10%	0.89
2	7	10	19%	5.30%	0.42
3	16	15	27.80%	10.40%	0.83
4	4	8	55.60%	15.90%	1.27
5	7	13	39.60%	15.10%	1.21
6	8	14	52.40%	22.80%	1.82

Table 2: Group characteristics

Finally, some of the most prevalent characteristics for the behavior of participants within the groups are presented below.

Group 1: The "evening at home" eaters: Group 1 holds the highest number of participants and through the analysis of the group most-significant rules and the actual triggering statistics, it was found that most participants in the group triggered rules when they were at "home" and especially during "evening" hours. Snacking at home could summarize the profile of this group.

Group 2: The "outdoors-social" eaters: Group 2 (already mentioned as the most healthy eating group) features among the most significant rules cases that involve "outdoors" or "other" as locations and "socializing" as circumstances. This comes in agreement with the "healthy-eating" assumption because it supports the fact that these participants eat unhealthy only in cases when they are out (e.g. in a restaurant, bar, etc.) and/or in the presence of others (which acts as a social influence factor as well).

Group 3: The "circumstances-driven" eaters: Group 3 features the highest number of rules (15) meaning that behavior within the group is more diverse (and also based on more complex rules). Analysis of triggered rules reveals that there are specific combinations of circumstances and locations that trigger most of the rules. Some of these combinations are: "ComputerRelated/Working and Home", "Traveling and Outdoors", "Other and Socializing".

Group 4: The "very-occasional" eaters: Group 4 is the group with the smallest number of participants and is considered to be a group that gathers participants that do not fit well with any of the other groups. It features very specific rules, applicable to other groups as well but in this case they are more prevalent, e.g. the rule that covers circumstances like “ComputerRelated” and “WatchingTV”, high positive emotions and unhealthy craving.

Group 5: The “after-activity” snackers: Group 5 has the main quirk characteristic that unhealthy eating is a result of either healthy cravings (or not cravings at all). Looking closely to the rule triggers revealed that activities within house or work (“HighLevelIn, LowLevel”) or “traveling” moments lead to unhealthy snacking despite the not-unhealthy cravings.

Group 6: The “unhealthy-cravings satisfaction” eaters: Group 6 features significant rules which are governed by the presence of unhealthy cravings that lead to unhealthy eating. Regardless emotions and time of day, these participants tend to indulge to their unhealthy cravings in various locations and under different circumstances. Not surprisingly, this is the group with the most triggers per day (almost 2).

5. Conclusion

In this paper, a framework for providing people with feedback regarding possible unhealthy eating events was presented. Data and analyses are based on a mobile application called “ThinkSlim” which was developed for the study, although the algorithm for building decision trees and extracting rules is generic and applicable to other datasets as well. Clustering of user rule activation vectors leads to six groups describing different eating patterns and can be used to build profiles that lead to unhealthy eating.

Further work involves more thorough analysis of the groups created so as to more precisely determine the characteristics of each group in regard to eating behavior. Moreover, a new study involving new participants using the “ThinkSlim” application with the above implemented framework is ongoing. Data from this study will be used to confirm the correctness of the approach and the validate the group description (profiling) with more data.

Acknowledgement

This study is funded by grant 12028 from Stichting Technische Wetenschappen (STW), Nationaal Initiatief Hersenen en Cognitie (NIHC), Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) and Philips under the Partnership programme Healthy Lifestyle Solutions.

References

1. B. A. Swinburn, G. Sacks, K. D. Hall, K. McPherson, D. T. Finegood, M. L. Moodie, S. L. Gortmaker, The global obesity pandemic: shaped by global drivers and local environments, The Lancet 378 (9793) (2011) 804–814.
2. M. J. Franz, J. J. VanWormer, A. L. Crain, J. L. Boucher, T. Histon, W. Caplan, J. D. Bowman, N. P. Pronk, Weight-loss outcomes: a systematic review and meta-analysis of weight-loss clinical trials with a minimum 1-year follow-up, Journal of the American Dietetic Association 107 (10) (2007) 1755–1767.
3. K. E. Heron, J. M. Smyth, Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments, British journal of health psychology 15 (1) (2010) 1–39.
4. C. N. Scollon, C.-K. Prieto, E. Diener, Experience sampling: promises and pitfalls, strength and weaknesses, in: Assessing well-being, Springer, 2009, pp. 157–180.
5. J. Zhou, F. Wang, J. Hu, J. Ye, From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2014, pp. 135–144.
6. W. Adler, S. Potapov, B. Lausen, Classification of repeated measurements data using tree-based ensemble methods, Computational Statistics 26 (2) (2011) 355–369.
7. R. J. Sela, J. S. Simonoff, RE-EM trees: a data mining approach for longitudinal and clustered data, Machine learning 86 (2) (2012) 169–207.
8. W.-Y. Loh, W. Zheng, et al., Regression trees for longitudinal and multiresponse data, The Annals of Applied Statistics 7 (1) (2013) 495–522.
9. W. Fu, J. S. Simonoff, Unbiased regression trees for longitudinal and clustered data, Computational Statistics & Data Analysis 88 (2015) 53–74.
10. A. A. Atienza, A. C. King, B. M. Oliveira, D. K. Ahn, C. D. Gardner, Using hand-held computer technologies to improve dietary intake, American journal of preventive medicine 34 (6) (2008) 514–518.
11. G. Spanakis, G. Weiss, B. Boh, A. Roefs, Network analysis of ecological momentary assessment data for monitoring and understanding eating behavior, in: Smart Health, Springer, 2015, pp. 43–54.
12. W.-Y. Loh, Y.-S. Shih, Split selection methods for classification trees, Statistica sinica (1997) 815–840.
13. O. Maimon, L. Rokach, Data mining and knowledge discovery handbook, Vol. 2, Springer, 2005.
14. B. Boh, L. H. Lemmens, A. Jansen, C. Nederkoorn, V. Kerkhofs, G. Spanakis, G. Weiss, A. Roefs, An ecological momentary intervention for weight loss and healthy eating via smartphone and internet: study protocol for a randomised controlled trial, Trials 17 (1) (2016) 1.
15. L. Kaufman, P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis, Vol. 344, John Wiley & Sons, 2009.