COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 137 (2016) 203–213

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Classifying smoking urges via machine learning

Antoine Dumortier [a], *Ellen Beckjord* [b], *Saul Shiffman* [c], *Ervin Sejdić* [a,*]

[a] *Department of Electrical and Computer Engineering, University of Pittsburgh, Benedum Hall, Pittsburgh, PA 15260, USA*

[b] *Department of Psychiatry, University of Pittsburgh, 5115 Centre Avenue, Suite 140, Pittsburgh, PA 15232, USA*

[c] *Department of Psychology, University of Pittsburgh, 510 BELPB, 130 N. Bellefield Avenue, Pittsburgh, PA 15260, USA*

ARTICLE INFO

Article history:
Received 14 May 2016
Received in revised form
12 August 2016
Accepted 20 September 2016

Keywords:
Smoking urges
Smoking cessation
Machine learning
Supervised learning
Feature selection

ABSTRACT

Background and objective: Smoking is the largest preventable cause of death and diseases in the developed world, and advances in modern electronics and machine learning can help us deliver real-time intervention to smokers in novel ways. In this paper, we examine different machine learning approaches to use situational features associated with having or not having urges to smoke during a quit attempt in order to accurately classify high-urge states.

Methods: To test our machine learning approaches, specifically, Bayes, discriminant analysis and decision tree learning methods, we used a dataset collected from over 300 participants who had initiated a quit attempt. The three classification approaches are evaluated observing sensitivity, specificity, accuracy and precision.

Results: The outcome of the analysis showed that algorithms based on feature selection make it possible to obtain high classification rates with only a few features selected from the entire dataset. The classification tree method outperformed the naive Bayes and discriminant analysis methods, with an accuracy of the classifications up to 86%. These numbers suggest that machine learning may be a suitable approach to deal with smoking cessation matters, and to predict smoking urges, outlining a potential use for mobile health applications.

Conclusions: In conclusion, machine learning classifiers can help identify smoking situations, and the search for the best features and classifier parameters significantly improves the algorithms' performance. In addition, this study also supports the usefulness of new technologies in improving the effect of smoking cessation interventions, the management of time and patients by therapists, and thus the optimization of available health care resources. Future studies should focus on providing more adaptive and personalized support to people who really need it, in a minimum amount of time by developing novel expert systems capable of delivering real-time interventions.

* *Corresponding author.* Department of Electrical and Computer Engineering, University of Pittsburgh, Benedum Hall, Pittsburgh, PA 15260, USA. Fax: +1-412-624-8003.
E-mail address: esejdic@ieee.org (E. Sejdić).

http://dx.doi.org/10.1016/j.cmpb.2016.09.016
0169-2607/

1. Introduction

There are 1.1 billion smokers in the world (15.4% of the world population), and this number is expected to increase to 1.6 billion over the next two decades [1]. Worldwide, tobacco use causes more than 5 million deaths per year [1]; that is to say, one person dies every six seconds from a tobacco related disease. Furthermore, current trends show that it will lead to the death of more than 8 million people annually by 2030 [1]. In the developed world, tobacco is currently the single largest preventable cause of death and diseases, and the overall mortality among both male and female smokers is about three times higher than that among similar people who never smoked [2]. In the United States, cigarette smoking kills more than 480,000 Americans each year, with more than 41,000 of these deaths from exposure to secondhand smoke. As a result, the economic consequences for the country are critical. Smoking-related illness in the United States costs more than $289 billion a year, including at least $133 billion in direct medical care for adults and $156 billion in lost productivity [2]. In 2012, an estimated 18.1% (42.1 million) U.S. adults were current cigarette smokers and 78.4% (33 million) of these adults smoked every day [3]. However, in 2011, 68.8% of adult cigarette smokers wanted to stop smoking completely [4], and 42.7% had made a quit attempt in the past year [2]. Unfortunately, quit attempts are typically unsuccessful, ending in relapse (resumption of smoking), usually precipitated by moments of intense craving or urge to smoke [5].

Studies based on ecological momentary assessment (EMA) [6,7] have been conducted. EMA involves repeated sampling of subjects' real world mood, thoughts and state of mind at specific and random times during the day, through completion of assessments in subject's daily routine using mobile technology [7,8]. In Refs. [9, 10], a subset of features previously associated with lapses is used to analyze how craving, emotion, and social environment impact on smoking rate [9]. In Ref. [10], self-reports of contextual variables were analyzed to examine correlates of craving when cigarettes were smoked. Results showed, for example, that craving was higher when cigarettes were smoked while eating or drinking, during activity, and early in the day. On the other hand, craving does not appear to be related with the location, alcohol, or caffeine. However, there is variability in the evidence regarding the degree to which different contextual features are associated with smoking risk. For example, during a quit attempt [11], self-reported temptation episodes (i.e., intense craving to smoke) were associated with negative moods, exposure to others smoking, and consumption of food, coffee, or alcohol.

These previous studies, as well as the recent advances in computational algorithms, has led us to believe that machine learning approaches can be useful in the process of smoking cessation. To test our hypothesis, we developed an algorithm used to predict subjects smoking urges, with a focus on the development of a general static algorithm intended for preliminary clinical testing. Our first goal was to carry out a comparative analysis of machine learning algorithms to determine if a classifier can be able to provide smoking urges' classifications. Then, implementing a feature selection algorithm, the second goal was to extract the most relevant features (in a given dataset) that can provide the best classifications of urges to smoke.

Section 2 describes the methodology for the different phases depicted. Before giving a presentation and technical information about the three classification methods used in this study, an explanation of how the input data selection is made and how data are organized is given. Validation techniques, which are implemented to split the dataset between a training dataset, to create the model, and a testing dataset, to test the model, are then presented. We also operate and compare feature selection algorithms in an attempt to exclude useless features and only select those which provide the best results. Section 3 presents the classification results and the final selected features, and Section 4 overviews the implications of these results. A conclusion is provided in Section 5 followed by a list of references.

2. Methodology

2.1. Data collection

The data were collected at the University of Pittsburgh from 1990 to 1995. A total of 349 smokers seeking to quit smoking were recruited through a media advertising and participants reported their smoking behavior, urges to smoke, and contextual information (e.g., mood, location) using a hand-held device at scheduled and random intervals five times a day for up to six weeks. The final dataset included 41 parameters (also called features) (Appendix A) and 29,959 environment reports from 248 unique subjects. The smoking urges are evaluated according to the value of one discrete attribute, which represents the urge rating at any point in time. This variable has its values on a 0 to 10 scale: 0 is for the lowest smoking urge, while 10 is for an intense smoking urge. In order to simplify the classification, the urge rating variable has been converted to a binary number. 0 (negative cases) has been attributed to values less than 5, and 1 (positive cases) has been attributed to values greater than or equal to 5. Out of these 29,959 reports, 70% (21,070 exactly) of them are 0, while 30% (8889 exactly) of the reports are 1. Our objective was to identify features associated with high urges (i.e., urge rating greater than or equal to 5).

2.2. Classification of smoking urges

Our unique dataset can be represented by the following form:

$$(\mathbf{X}, \mathbf{Y}) = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots, \mathbf{X}_k, \ldots, \mathbf{X}_n, \mathbf{Y}) \quad (1)$$

where $\mathbf{Y}$ (binary column vector) is the target variable (the class) of the predictions (the urge rating). The matrix $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots, \mathbf{X}_n)$ represents the features used by the classifier (each $\mathbf{X}_k$ is a column of $\mathbf{X}$). $\mathbf{X}$ can also be represented using the subject approach, $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n)^T$ where each $\mathbf{x}_k$ is a line of $\mathbf{X}$. Therefore, we have: