Reality mining in eHealth

Peter Wlodarczak

wlodarczak@gmail.com

Prof. Jeffrey Soar
Jeffrey.Soar@usq.edu.au

Dr. Mustafa Ally
Mustafa.Ally@usq.edu.au

Abstract. There is increasing interest in Big Data analytics in health care. Behavioral health analytics is a care management technology that aims to improve the quality of care and reduce health care costs based capture and analysis of data on patient's behavioral patterns. Big Data analytics of behavioral health data offers the potential of more precise and personalized treatment as well as monitor population-wide events such as epidemics.

Mobile phones are powerful social sensors that are usually physically close to users and leave digital traces of users' behaviors and movement patterns. New Apps (application or piece of software) are emerging that passively collect and analyze mobile phone data of at-risk patients such as their location, calling and texting records and app usage, and can find deviations in a user's daily patterns to detect that something is wrong before an event occurs. Data mining and machine learning techniques are adopted to analyze the "automated diaries" created by the smart phone and monitor the well-being of people. The App first learns a patients daily behavioral patterns using machine learning techniques. Once trained, the App detects deviations and alerts carers based on predictive models.

This paper describes the techniques used and algorithms for reality mining and predictive analysis used in eHealth Apps.

Keywords: Reality mining, Big Data, machine learning, eHealth, predictive analytics, behavioral health analytics, mobile sensing

Reality mining in eHealth, p. 1, 2014.

1 Introduction

An important question for behavioral epidemiology and public health is to better understand how individual behavior is affected by illness and stress [3]. Someone who becomes depressed isolates himself and has a hard time to get up and go to work. He shows deviations from his normal behavioral patterns. Smartphones produce significant amounts of behavioral data. They are essentially off-the-shelf wearable computers. They can provide a convenient tool for measuring social connectivity features related to phone calls and text messages [1]. Users usually keep mobile phones physically close to themselves. The mobile sensor data thus reflects the same movement patterns as the user. Most Smartphones are equipped with accelerometers for motion detection, GPS (Global positioning system) monitor where a user visits and call logs record call duration. They are powerful social sensors for spatio-temporal data. Decreased movement detected by motion sensors or infrequent texts in the message log might be symptoms of depression. Shorter than usual calls might signal isolation.

Real-time data collection and analysis of mobile phone data reveals information on the health state of a user and can be used to diagnose if a patient becomes symptomatic and prompt early treatment. Symptoms that can be detected are anxiety, stress, disease spread, and obesity [2, 3]. If symptoms are detected, a health care center can be alerted and a nurse can call the patient and check on his situation. This type of proactive healthcare is especially useful for high risk patients or patients susceptible of underreporting like mentally ill or elderly people.

2 Methodology

Reality mining refers to the process of collecting and analyzing machine sensed human behavioral data such as movement patterns, human interactions and human communication patterns, with the goal of detecting predictable behavioral patterns [18]. Reality mining comprises four phases. A data collection phase, a data pre-processing phase, a data mining phase and a post-processing phase. Sometimes a predictive analysis phase is added. Here the predictive step is considered part of the post-processing.

2.1 Data collection

The data collection phase records a patient’s behavioral data from interactions from electronic exchanges (call records, SMS logs, email headers) and contextual data (location information). Sometimes other data like face-to-face proximity for individuals has been collected too using the mobile phones Bluetooth connection [2]. The mobile phone is used to extract conversational partners and location of a user, that is, the total number of interactions, the diversity of interactions, and the diversity (entropy) of his behavior [2]. Smartphones provide APIs to access the underlying functionality such as GPS sensors or call logs programmatically.

2.2 Data pre-processing

Not all data collected is useful. The data has thus to be relevance filtered first. Also the raw sensor data is not in a format that can be used by most ML algorithms. The data has to be transformed into a feature vector. A feature in a feature vector can represent the coordinates of a location or the call duration. Eigenvector analysis, commonly known as principal components analysis, is the optimal linear method for obtaining a low-dimensional approximation to a signal such as observations of user behavior [5]. Behavioral structure can be represented by the principal components of the spatiotemporal data set, termed eigen-behaviors [10]. The term eigenbehavior was introduced by Eagle and Pentland [11]. We represent this behavioral structure by the principal components of the complete behavioral dataset, a set of characteristic vectors we have termed eigenbehaviors [11]. Eigenbehaviors provide an efficient data structure for learning and classifying tasks.

To calculate the Eigenbehavior a person's behavior has to be measured, for instance the time sequence of their phone calls or text messages. For a group of *M* people, and the behaviors Γ_1, Γ_2, …, Γ_M, the average behavior is:

$$\psi = \frac{1}{M}\sum_{n=1}^{M}\Gamma_{\mathrm{n}} \tag{1}$$

A set of *M* vectors, $\Phi_i = \Gamma_i - \Psi$, is defined to be the deviation of the normal behavior. Principle components analysis is subsequently performed on these vectors generating a set *M* orthonormal vectors, u_n, which best describes the distribution of the set of behavior data when linearly combined with their respective scalar values, λ_n [5]. The Eigenvector and Eigenvalues of the covariance matrix of Φ are calculated as:

$$C = \frac{1}{M}\cdot\sum_{n=1}^{M}\phi_n\cdot\phi_n^T = A\cdot A^T \tag{2}$$

Where the Matrix $A = [\Phi_1, \Phi_2, \ldots, \Phi_M]$. A typical daily pattern is leaving the sleeping place in the morning, spending time in a small set of locations during office hours, and occasionally moving to a few locations in the evening and on the weekends. For typical individuals the top three Eigenbehavior components account for up to 96% of the variance in their behavior [5]. This means that a person's location context can be classified with high accuracy.

2.3 Data mining

To make predictions about a person's health state, the behavioral data needs to be automatically classified into normal and deviant behavior. Machine learning (ML) techniques have been successfully applied for data classification problems. A given data set is typically divided into two parts: training and testing data sets with known class labels [8]. The class label is "normal" and "deviant" behaviour. The training data is the data collected during a training phase to learn a patient's normal behaviour as represented by the Eigenvector. It is used to train a model. The test data, the real-time behavioural data, is then applied against the trained model. The model analyses the data for any abnormalities and makes predictions about the health state. Typical supervised learning methods include naïve Bayes classification, decision tree induction, k-nearest neighbors, and support vector machines [4]. There are many more ML algorithms. Experience shows that no single machine learning scheme is appropriate to all data mining problems [9]. Usually several algorithms are trained and compared to determine which one gives the most accurate results for a given problem [6].

Ultimately we want to obtain a decision function *f*, that classifies the behavioral pattern *h* as normal (*N*), or deviant (*D*). If we denote the set of all behavioral patterns by *H*, we search for a function $f{:}H \rightarrow \{N,D\}$. We use the set of behavioral data collected during the training phase $\{(h_1, c_1), (h_2, c_2), \ldots, (h_n, c_n)\}$, where: $h_i \in H$, $c_i \in \{N, D\}$, to train the model. The naïve Bayes classifier is a family of simple probabilistic classifiers based on the Bayes theorem [6]. Decision tree learning creates decision trees, where a decision could be: did the patient go to coordinate x,y early in the morning, yes/no. Support Vector Machine (SVM) classifications are based on statistical learning theory and classifies data by separating them with a hyperplane. Which classification algorithm performs best can depend on the type of illness, but other factors such as the patient's normal behavior have an influence on the accuracy. Once the model is trained, it can be used for predictions based on real-time data collected through the mobile phone. ML techniques are well documented in literature [6,9,12] and are not further explored here.

2.4 Data post-processing

Characteristic behavioral changes can be associated with symptoms based on the classification scheme from behavioral patterns. In the Susceptible, Infectious, Recovered or SIR model especially in the S(usceptible) to I(nfectious) transition phase user behavior changes [3] and can thus be used to improve prediction accuracy.

To analyze the temporal relationship of the behavior, Granger causality analysis has been used. The traditional linear Granger test has been widely used to examine the linear causality among several time series in bivariate settings as well as multivariate settings [14]. It is used to determine if one time series has predictive information for another. For a behavioral pattern a time series can be for instance the coordinates of places a patient frequently visited during the training phase, for instance the coordinates of his work place or favorite café. The second time series are the coordinates of places he visits over time during the testing phase.

The original Granger tests examined the linear causality among several time series in a bivariate and multivariate setting. However many real world applications are nonlinear and extensions have been developed [14, 16] to overcome this constraint. Recently the Phase Slope Index (PSI) has been preferred over Granger causality in some studies [3], [17]. PSI is a recently proposed spectral estimation method designed to measure temporal information flux between time series signals [3]. It is based on the assumption that the information flux between two signals can be estimated using the phase slope of the cross-spectrum of the signals. Independent noise mixing does not affect the complex part of the coherency between multivariate spectra, and hence PSI is considered more noise immune than Granger analysis [3]. The Phase Slope Index is defined as:

$$\Psi_{ij} = \Im(\sum_{f \in F} C_{ij}^{*}(f) C_{ij}(f + \delta f)) \tag{3}$$

Where C_{ij} is the complex coherency, $\delta f = 1/T$ is the frequency resolution, and $\Im(\cdot)$ denotes taking the imaginary part. PSI has been used to validate causal links between time series of symptom days where participants showed stress and depression symptoms [2].

3 Challenges and ethical issues

Behavioral patterns are highly personal and vary from individual to individual. Behavioral patterns of introverts, persons lacking social skills, lethargic or isolated persons show smaller variations when sick than active, sociable persons. Training and predictive models have to be enough granular to capture and detect deviant behavior of patients with a big variety of different behavioral patterns. There are many reasons why behavioral patterns change. Students before examinations spend more time studying and are less engaged in physical activities. Someone in a new relationship might change his behavioral patterns. The challenge of correctly classifying behavior and avoiding false positives based on misinterpretation like “work at home” interpreted as deviant behavior has to be addressed by any real-world application. Recording and analyzing the behavioral patterns of patients in real-time raises serious privacy issues. It represents a high level of surveillance where every movement and conversation is logged for analysis. There are also security issues. Announcing a person’s location to the world can tip off burglars or stalkers.

4 Conclusions

While reality mining on mobile phones in the health care sector is still in its infancy, there are already promising applications. Modern societies face the challenge of caring for their aging population. Applications of reality mining using mobile phones might help elderly people, people with disabilities or diseases like Alzheimer’s living safer and more independently and reduce health care costs. But there seems to be no boundary for further applications on the individual level as well as on the public health level. Reality mining has already been used to measure social interactions or movement patterns of populations to determine the spread of infectious diseases and studies have buttressed the effectiveness of cell phones for early detection of outbreaks of epidemics [1,2]. There are already projects studying the spread of diseases in Africa [7]. Our findings suggest that it might be possible to answer such questions in the near future and to begin planning how to influence the development of even greater health-sensing capabilities in smartphones [2].

Lastly reality mining has shown that humans are more predictable than believed and that it is thus possible to reveal the identity of a person even if the mobile phone data is anonymized. More research in anonymizing behavioral patterns in reality mining would be highly desirable especially when used in the eHealth area.

5 References

1. I. Chronis, A. Madan, and A. Pentland, "SocialCircuits: the art of using mobile phones for modeling personal interactions," in Proceedings of the ICMI-MLMI '09 Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing, Cambridge, Massachusetts, 2009, pp. 1-4.
2. A. Madan, M. Cebrian, S. Moturu, K. Farrahi, and A. Pentland, "Sensing the "Health State" of a Community," *Pervasive Computing, IEEE,* vol. 11, no. 4, pp. 36-45, 2012.
3. A. Madan, M. Cebrian, D. Lazer, and A. Pentland, "Social sensing for epidemiological behavior change," in Proceedings of the 12th ACM international conference on Ubiquitous computing, Copenhagen, Denmark, 2010, pp. 291-300.
4. P. Gundecha, and H. Liu, "Mining Social Media: A Brief Introduction," informs, vol. 9, pp. 1-17, 2012.
5. A. Pentland, "Automatic mapping and modeling of human networks," Physica A: Statistical Mechanics and its Applications, vol. 378, no. 1, pp. 59-67, 5/1/, 2007.
6. P. Wlodarczak, J. Soar, and M. Ally, "What the future holds for Social Media data analysis," World Academy of Science, Engineering and Technology, vol. 9, no. 1, pp. 545, 2015.
7. "Big Data Gets Personal," MIT Technology Review, vol. 116, no. 4, 2013.
8. K. Tretyakov, "Machine Learning Techniques in Spam Filtering," Data Mining Problem-oriented Seminar, U. o. T. Institute of Computer Science, ed., 2004, p. 19.
9. I. H. Witten, E. Frank, and M. A. Hall, Data Mining, 3 ed., Burlington, MA, USA: Elsevier, 2011.
10. K. Sookhanaphibarn, R. Thawonmas, F. Rinaldo, and K.-T. Chen, "Spatiotemporal analysis in virtual environments using eigenbehaviors," in Proceedings of the 7th International Conference on Advances in Computer Entertainment Technology, Taipei, Taiwan, 2010, pp. 62-65.
11. N. Eagle, and A. Pentland, "Eigenbehaviors: identifying structure in routine," Behavioral Ecology and Sociobiology, vol. 63, no. 7, pp. 1057-1066, 2009/05/01, 2009.
12. B. Liu, Sentiment Analysis and Opinion Mining: Morgan & Claypool, 2012.
13. B. Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2 ed., Heidelberg: Springer, 2011.
14. Z. Bai, W.-K. Wong, and B. Zhang, "Multivariate linear and nonlinear causality tests," Mathematics and Computers in Simulation, vol. 81, no. 1, pp. 5-17, 9//, 2010.
15. C. Diks, and V. Panchenko, "A new statistic and practical guidelines for nonparametric Granger causality testing," Journal of Economic Dynamics and Control, vol. 30, no. 9–10, pp. 1647-1669, 9//, 2006.
16. C. Hiemstra, and J. D. Jones, "Testing for Linear and Nonlinear Granger Causality in the Stock Price- Volume Relation," The Journal of Finance, vol. 49, no. 5, pp. 1639-1664, 1994.
17. G. Nolte, A. Ziehe, N. Krämer, F. Poupescu, and K.-R. Müller, "Comparison of Granger Causality and Phase Slope Index," in NIPS08 workshop on Causality, Canada, 2008.
18. T. Simonite, "Smartphone Tracker Gives Doctors Remote Viewing Powers," Technology review, vol. 116, no. 4, 2013.