EAI Endorsed Transactions
on Scalable Information Systems

Research Article ICST.ORG

Behavioural health analytics using mobile phones

P. Wlodarczak[1],*, J. Soar[1] and M. Ally[1]

[1] University of Southern Queensland, Toowoomba, Queensland, Australia

Abstract

Big Data analytics in healthcare has become a very active area of research since it promises to reduce costs and to improve health care quality. Behavioural analytics analyses a patients behavioural patterns with the goal of early detection if a patient becomes symptomatic and triggering treatment even before a disease outbreak happens. Behavioural analytics allows a more precise and personalised treatment and can even monitor whole populations for events such as epidemic outbreaks.

With the prevalence of mobile phones, they have been used to monitor the health of patients by analysing their behavioural and movement patterns. Cell phones are always on devices and are usually close to their users. As such they can be used as social sensors to create "automated diaries" of their users. Specialised apps passively collect and analyse user data to detect if a patient shows some deviant behaviour indicating he has become symptomatic. These apps first learn a patients normal daily patterns and alert a health care centre if it detects a deviant behaviour. The health care centre can then call the patient and check on his well-being. These apps use machine learning techniques to for reality mining and predictive analysis. This paper describes some of these techniques that have been adopted recently in eHealth apps.

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

Received on 30 June 2015, accepted on 30 June 2015, published on 02 July 2015

Copyright © 2015 P. Wlodarczak *et al.*, licensed to ICST. This is an open access article distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/3.0/), which permits unlimited use, distribution and reproduction in any medium so long as the original work is properly cited.

doi: 10.4108/sis.2.5.e6

1. Introduction

An important question in behavioural epidemiology and public health is to understand how individual behaviour 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 or she shows deviations from his normal behavioural patterns. Smartphones produce a significant amount of behavioural data. They are essentially off-the-shelf wearable computers. They provide a convenient tool for measuring social connectivity features related to phone calls and text messages [1]. Users usually keep mobile phones close to themselves and, unlike land line phones, are generally used by just one person. 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) for monitoring where a user visits and call logs for recording 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 centre can be alerted and a nurse can call the patient and check on their situation. This type of proactive healthcare is especially useful for

*Corresponding author. Email:wlodarczak@gmail.com high risk patients or patients susceptible of underreporting like the mentally ill or the elderly.

A critical requirement for the ability to predict if a patient becomes symptomatic is the capability to learn and categorize their behaviour. A common way of analysing behavioural data and make predictions is by applying machine learning techniques. A model is trained using behavioural data. It first learns the user's normal behaviour for example where he goes, who he calls and texts, call durations and email traffic. In this reality mining phase the data is collected and used to train a machine learning algorithm such as the naïve Bayes classifier or k-Nearest Neighbor [4]. Once trained the model can be used for predictive modelling of individual symptoms based on deviant user behaviour that has been detected.

The next section describes the reality mining and predictive analysis steps in more detail.

2. Methodology

Reality mining and predictive analysis 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. Some of the steps are iterative and are not necessarily processed sequentially. Figure 1 shows the mining and analysis steps.

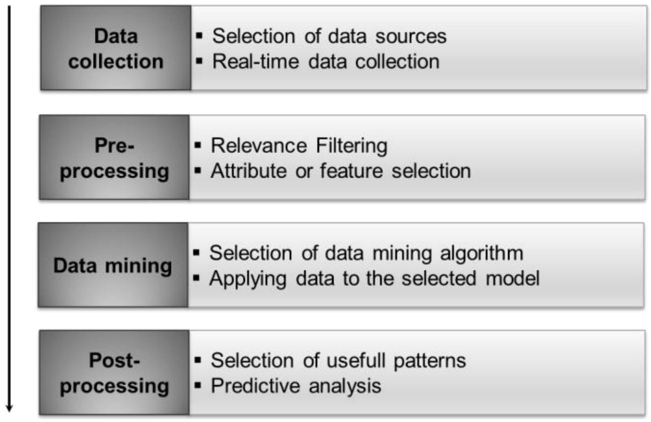
Data collection
▪ Selection of data sources
▪ Real-time data collection
Pre-processing
▪ Relevance Filtering
▪ Attribute or feature selection
Data mining
▪ Selection of data mining algorithm
▪ Applying data to the selected model
Post-processing
▪ Selection of usefull patterns
▪ Predictive analysis

Figure 1. Reality mining phases

2.1. Data collection

The data collection phase records a patient's behavioural data from interactions from electronic exchanges (call records, SLS logs, and email headers) and contextual data (location information). Sometimes other data like face-to-face proximity for individuals has been collected as well using the mobile phone's 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 behaviour [2].

Android based Smartphones and the iPhone provide APIs (Application Programming Interfaces) for accessing mobile phone sensor data. They can be accessed in real-time and stored on the device. Usually on Android based Smartphones Apps are developed using the Java programming language, and on iPhone Apps using Objective C. But other programming languages like ANSI C can be used instead.

The methods in Figure 2 register an Android App with the GPS Location Service and poll every 3 seconds for the longitude and latitude coordinates:

```
@Override
protected void onCreate(Bundle
savedInstanceState) {
   super.onCreate(savedInstanceState);

setContentView(R.layout.activity_gps_bas
ics);
   locationManager = (LocationManager)
getSystemService(Context.LOCATION_SERVIC
E);

locationManager.requestLocationUpdates(
LocationManager.GPS_PROVIDER, 3000, 10,
this);
}

@Override
public void onLocationChanged(Location
location) {
   String str = "Latitude:
"+location.getLatitude()+"Longitude:
"+location.getLongitude();
   Toast.makeText(getBaseContext(), str,
Toast.LENGTH_LONG).show();
}
```

Figure 2. Java code to get GPS data

2.2. Data pre-processing

Not all data collected is useful. For instance if someone goes to the same café every morning only the location data of the café might be of interest, not the movement data to get there. The data has thus to be relevance filtered first. Data purification is an important pre-processing step in reality mining. Bitter experience shows that real data is often disappointingly low in quality, and careful checking - a process that has become known as data cleaning - pays off many times over [9]. Many Machine Language (ML) algorithms handle noise poorly and the accuracy of the prediction is impaired if the raw data is not first properly pre-processed.

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 be the coordinates of a location or a 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 behaviour [5]. Behavioural structure can be represented by the principal components of the spatiotemporal data set, termed eigenbehaviors [10]. The term eigenbehavior was introduced by Eagle and Pentland [11]. We represent this behavioural structure by the principal components of the complete behavioural 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 behaviour has to be measured, for instance the time sequence of their phone calls or text messages. For a group of *M* people, and the behaviours $\Gamma_1, \Gamma_2, \ldots, \Gamma_M$, the average behaviour 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 behaviour [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 behavioural data needs to be automatically classified into normal and deviant behaviour. Often Machine Learning (ML) techniques are used for classification. It should be noted that many classifiers output the probability *Pr*, that a behaviour x_i with corresponding label y_i belongs to class *j*[17]:

$$\Pr(x_i | y_i = j) \tag{3}$$

ML techniques are divided into supervised, unsupervised and semi-supervised methods. For classification, supervised methods are used. They are adopted when the class label is known. Here the class label is "normal" and "deviant" behavior. Unsupervised methods are used when the class label is unknown and semi-supervised methods are used when there are small amounts of data where the class label is unknown and large amounts where it is unknown. Unsupervised methods have been used for classification in some cases, for instance in sentiment analysis [12].

For supervised learning algorithms, a given data set is typically divided into two parts: training and testing data sets with known class labels [8]. Here 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 for all data mining problems [9]. That is why usually several algorithms are trained and compared to determine which one has the highest classification accuracy. The classification accuracy *A* is defined as:

$$A = \frac{C}{T} \tag{4}$$

Where *C* is the number of correct classifications and *T* the total number of test cases. 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. 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. Naïve Bayes, decision tree and SVM learn a model from training data, posterior probabilities or sets of rules. They generalize. They are called eager learning. k-nearest neighbor (k-NN) in contrast is a lazy learning method where no model is learned. Learning only occurs when a test example needs to be classified [13]. k-NN is a non-parametric method that takes the k closest training example for classification. $k = 1$ is not sufficient for classification and a set of nearest neighbors is used.

There are various approaches to determine the best performing algorithm for a specific problem. The simplest approach is called 1-ErrorRate which corresponds to the classification accuracy.

A more sophisticated method is n-fold cross-validation where *n* is usually between 5 and 10. In cross-validation, you decide on a fixed number of folds, or partitions, of the data [9]. If *n* is 10, *n* – 1 folds are used for training, 1 for testing. The data is randomly divided into 10 folds. Each fold is held out in turn and the learning scheme is trained using the remaining *n* – 1 folds and the error rate is calculated on the hold out fold. This process is executed 10 times on different training sets. Finally the error is averaged on the 10 error estimates. This way of predicting the error rate is called *stratified tenfold cross-validation* and is a standard learning technique.

Which classification algorithm performs best can depend on the type of illness, but other factors, like the patient's normal behavior, have an influence on the accuracy.

Data mining typically goes through many iterations until satisfactory results are achieved. Once the model is trained, it can be used for predictions based on real-time data collected through the mobile phone.

2.4. Data post-processing

Characteristic behavioral changes can be associated with symptoms based on the classification scheme from behavioral features. 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 about 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. For a strictly stationary bivariate process $\{(X_t, Y_t)\}$, $\{X_t\}$ is a Granger cause of $\{Y_t\}$ if past and current values of X contain additional information on future values of Y that is not contained in past and current Y-values alone [15]. The Granger causality test for two scalar-valued, stationary, and ergodic time series $\{X_t\}$ and $\{Y_t\}$ is defined as:

$$F(X_t \mid \mathrm{I}_{\text{t-1}}) = F(X_t \mid (\mathrm{I}_{\text{t-1}} - Y_{t-Ly}^{Ly})), t = 1,2... \qquad (5)$$

Where $F(X_t|I_{t\text{-}1})$ is the conditional probability distribution of X_t given the bivariate set $I_{t\text{-}1}$ consisting of an Lx-length vector X_t and an Ly-length vector of Y_t. If the equality in equation (4) does not hold, then knowledge of past Y values helps to predict current and future X values, and Y is said to strictly Granger cause X [16].

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], [6]. 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:

$$\Psi_{ij} = \Im(\sum_{f \in F} C_{ij}^{*}(f) C_{ij}(f + \delta f)) \qquad (6)$$

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 granular enough 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, for instance someone doing home office is interpreted as deviant behavior, has to be addressed by any real-world application.

Noise is a serious problem for any data mining application. ML algorithms usually handle noise poorly and reduce the predictive accuracy. Noise also affects the performance of the App.

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. The privacy might not only infringe on the cell phone user himself but on the persons he calls or are close to him. The feeling of being constantly under study or surveillance can make people feel uneasy.

There are also security issues. Announcing a person's location to the world can tip off burglars or stalkers.

Social acceptance is constantly changing and when the usefulness exceeds the detriments then social norms are likely to change and what is perceived as intrusion today might become acceptable in the future. In any event users should always be kept informed about what data is being collected and who has access to them.

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 challenges of caring for their aging populations. Applications of reality mining using mobile phones might help elderly people, people with disabilities or diseases like Alzheimer's to live safer and more independently and reduce health care costs. But there seems to be no boundaries 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 [2] 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].

Previous research on behavioral patterns and the health state of patients depended on surveys and experience sampling. Mobile phones, through their location and movement detection capabilities and call and text logs, allows us to quantify behavior with much more accuracy, higher resolution, and continuity.

However there are serious privacy concerns and future research needs to focus on the privacy aspects of data collection, analysis and preservation.

Also, 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.

While there is a great deal of research in the mining and analyzing of data, there seems to be less emphasis placed in the data purification process. Due to the vast amounts of data that can be collected using mobile phones, filtering the relevant data and reducing noise is where much optimization can be achieved. Also a mobile phone can be enhanced by attaching more sensors. Mobile phones are already used as electrocardiograms [7] but other sensors to measure blood pressure, stress levels or saliva composition can be attached. Future research could address these issues and make eHealth Apps more useful.

References

[1] Journal article: Cronis, I. Madan, A. and Pentland, A. (2009) SocialCircuits: the art of using mobile phones for modeling personal interactions. *Proceedings of the ICMI-MLMI '09 Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing*: 1-4.

[2] Journal article: Madan, A. Cebrian, M. Moturu, S. Farrahi, K. and Pentland A. (2012) Sensing the "Health State" of a Community," *Pervasive Computing, IEEE,* volume(11): 36-45.

[3] Journal article: Madan, A. Cebrian, M. Lazer, D. and Pentland, A. (2010) Social sensing for epidemiological behaviour change, *Proceedings of the 12th ACM international conference on Ubiquitous computing*: 291-300.

[4] Journal article: Gundecha, P. and Liu, H. (2012) Mining Social Media: A Brief Introduction, *informs*, volume(9): 1-17.

[5] Journal article: Pentland, A. (2007) Automatic mapping and modeling of human networks, *Physica A: Statistical Mechanics and its Applications*, volume(378): 59-67.

[6] Journal article: Nolte, G. Ziehe, A. Krämer, N. Poupescu, F. and Müller, K.-R. (2008) Comparison of Granger Causality and Phase Slope Index, *NIPS08 workshop on Causality*.

[7] Journal article: (2013) Big Data Gets Personal, *MIT Technology Review*, volume(116).

[8] Journal article: Tretyakov, K. (2004) Machine Learning Techniques in Spam Filtering, *Data Mining Problem-oriented Seminar*, p. 19.

[9] Book: Witten, I. H. Frank, E. and Hall, M. A. (2011) *Data Mining*, 3 ed. (Burlington, MA, USA: Elsevier).

[10] Journal article: Sookhanaphibarn, K. Thawonmas, R. Rinaldo, F. and Chen, K.-T. (2010) Spatiotemporal analysis in virtual environments using eigenbehaviors, *Proceedings of the 7th International Conference on Advances in Computer Entertainment Technology*, 62-65.

[11] Journal article: Eagle, N. and Pentland, A. (2009) Eigenbehaviors: identifying structure in routine, *Behavioral Ecology and Sociobiology*, volume(63), 1057-1066.

[12] Book: Liu, B. (2012) *Sentiment Analysis and Opinion Mining* (Morgan & Claypool).

[13] Book: Liu, B. (2011) *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, 2 ed. (Heidelberg: Springer).

[14] Journal article: Bai, Z. Wong, W.-K. and Zhang, B. (2010) Multivariate linear and nonlinear causality tests, *Mathematics and Computers in Simulation*, volume(81) 5-17.

[15] Journal article: Diks, C. and Panchenko, V. (2006) A new statistic and practical guidelines for nonparametric Granger causality testing, *Journal of Economic Dynamics and Control*, volume(30), 1647-1669.

[16] Journal article: Hiemstra, C. and Jones, J. D. (1994) Testing for Linear and Nonlinear Granger Causality in the Stock Price- Volume Relation, *The Journal of Finance*, volume(49), 1639-1664.

[17] Journal article: Wlodarczak, P. Soar J. and Ally, M. (2015) Genome mining using machine learning techniques, *Inclusive Smart Cities and e-Health*, volume(9102), 379-384.