

Medical diagnostics based on combination of sensor and user-provided data

Maja Somrak¹, Anton Gradišek¹, Mitja Luštrek¹, Ana Mlinar²,
Miha Sok³, and Matjaž Gams¹

¹ Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

² MESI d.o.o., Leskoškova cesta 9d, 1000 Ljubljana, Slovenia

³ Faculty of Medicine, University of Ljubljana, Vrazov trg 2, 1000 Ljubljana, Slovenia

{maja.somrak,mitja.lustrek}@ijs.si

Abstract. We present an approach that incorporates multiple machine-learning and data mining algorithms for prediction of the user's medical condition. The decisioning is based on vital signs data and user-provided input regarding the symptoms expressed. The presented method was trained and tested on virtual patients, generated using expert medical knowledge. We discuss future steps in the method development.

Keywords: medical diagnostics, questionnaire, sensor data, vital signs, user input

1 Introduction

Being able to diagnose common diseases at home can help the patient decide when to go to the doctor and at the same time reduce the burden on healthcare system. Here, we present a diagnostic method that incorporates information about user symptoms and vital signs readings to predict user's medical condition. The method was developed as a part of the diagnostic software for *HealthStation HOME*, a device competing at the Qualcomm Tricorder XPRIZE \$10 million challenge [1]. We present initial results of experimental tests on virtual patients and outline possible improvements in future.

1.1 HealthStation HOME

The HealthStation HOME system [2] consists of a set of sensors that measure vital signs, such as heart rate, breathing rate, blood pressure, body temperature and blood oxygen saturation. The collected data is communicated to a mobile device for further diagnostics. The measurements can be interpreted as pathological symptoms (e.g. high blood pressure) that serve as an input for the diagnostic application. The user obtains an evaluation of his health condition by running the diagnostic application on the mobile device by selecting one of the starting options, *I feel pain* or *I feel unwell* (see Fig. 1). The application then guides the user through intelligently selected questions about the symptoms that are recognized as relevant. The result of the diagnostic method is the initial, home-based,

medical condition assessment and a recommendation for further diagnostic testing (additional HealthStation HOME modules, such as blood or urine tests) to finally confirm or reject the diagnosis.

2 The Method

The HealthStation HOME diagnostic method is designed in form of a questionnaire with multi-modal inputs. The overview of the procedure is shown in Fig. 1. The initial input for the method consists of three types of data: (1) identified risk factors from the user profile data (e.g. smoking), (2) recognized pathological symptoms from vital signs measurements or other sensor data (e.g. high blood pressure) and (3) user selected pain symptoms in the application (e.g. chest pain). There are 60 predefined symptoms that the method can operate with, each of which can be treated as *unknown* or *known*, where *known* symptom is either *present* or *absent* in a patient. The *present* symptoms from the three inputs form the initial set of symptoms (4), which serve as the basis for automatically compiling a list of additional symptoms (5), from which the user is expected to select those he/she is experiencing. This list is generated to include both the symptoms that the user most probably experiences at the time and would probably want to report, and also the most relevant symptoms that would help the physician or the diagnostic method set a reliable diagnosis.

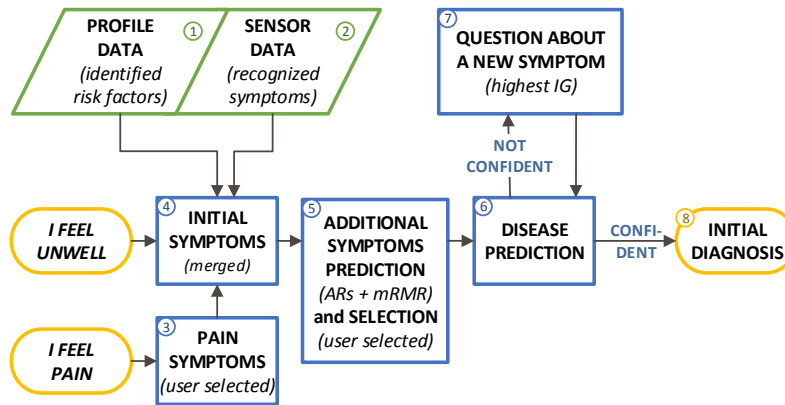


Fig. 1. Method overview. The numbers in circles indicate step numbers, as explained in the text.

If there is at least one symptom in the initial symptoms set, the method aims to find other symptoms that often emerge together with one of these initial

symptoms. For this purpose, association rules (ARs) [3–5] of type *symptom A* → *symptom B* are searched, where symptom A is any of the initial symptoms and symptom B is any of the unknown symptoms. Rules with the highest confidence and minimal support condition satisfied are selected in order to produce a set of probable additional symptoms.

In addition to the ARs, the minimum-Redundancy-Maximum-Relevance [6] (mRMR) method is used to identify the most informative, mutually independent symptoms that have not been examined yet (unknown symptoms). This method works even if there are no symptoms in the initial symptom set. The mRMR resulting attributes subset is the subset of attributes (symptoms) that a) provide a lot of information about the class (medical condition) and b) are at the same time mutually uncorrelated. The criterion a) is measured with the mutual information between each attribute and the class, while b) is measured with the mutual information between the attributes. The mRMR rule used in our method is defined upon mutual information difference (MID) criterion [6] with the following equation

$$\max_{i \in \Omega_S} [I(i, h) - \frac{1}{|S|} \sum_{j \in S} I(i, j)]. \quad (1)$$

For selecting each additional symptom, an iteration of mRMR calculation over all unknown symptoms is repeated to find a symptom for which the value of the function is maximized (Eq. 1). $I(i, h)$ is mutual information between i , a symptom from the set of unknown symptoms Ω_S , and h , the classification variable – the medical condition. Likewise, j is a symptom from the set of known symptoms S , containing $|S|$ symptoms. Once a new symptom i is selected and added to the additional set of symptoms, it is treated as one of the known symptoms (the symptom is moved from Ω_S to S) in next iteration of mRMR calculation.

In the following step, the information about present and absent symptoms is first used for the disease prediction (6). The probabilities for predefined 15 different medical conditions (14 diseases and 'healthy') are evaluated using a set of J48 classifiers, one for each condition. There are two probability thresholds that represent medium and high chance for a certain medical condition, empirically selected to be 40% and 80%, respectively. If all condition probabilities fall below the medium threshold (improbable condition) or above the high threshold (very probable condition), the prediction is considered confident and the diagnostic procedure terminates, retrieving the diagnosis. However, if one or more medical condition probabilities lie between the medium and high threshold (neither very probable nor improbable condition), in the so called *gray zone*, the disease prediction is not considered confident. In this case, information about at least one additional symptom is needed to obtain a confident prediction. This is obtained by asking user a question about a new symptom (7). The additional symptom is chosen according to the highest information gain (IG), where the values are recalculated from a reweighted training set, such that the instances with the conditions from the *gray zone* are assigned higher weights. This approach, espe-

cially when incorporating reweighting, reduces the number of questions required for the probabilities to emerge out of the *gray zone*, as opposed to randomly selected questions. In case any condition probabilities still remain in the *gray zone* after a maximum number of question has been asked, the procedure terminates, selecting the medical condition with the highest probability for the diagnosis.

3 Experiments

We utilized expert medical knowledge to obtain the patient data sets. For this purpose, a table correlating 15 different medical conditions with over 60 symptoms was developed by physicians. The table was used for generating the training set containing 15000 virtual patients. Additionally, a test set of 1500 virtual patients was generated, where each medical condition was present in 100 patients. The tests demonstrated high sensitivity and specificity. For example, for otitis media, 94% of the patients with this medical condition were correctly identified while 84% of patients, diagnosed with otitis media actually had the disease. In the case of leukocytosis, the corresponding values were 59% and 60%, respectively [7]. On average for all medical conditions, the obtained values for sensitivity were 88.4%, for specificity 88.6% and for accuracy 88.3%. Currently, we are collecting the data of real patients for further testing, an example of a patient answering to the questions is shown in Fig. 2.

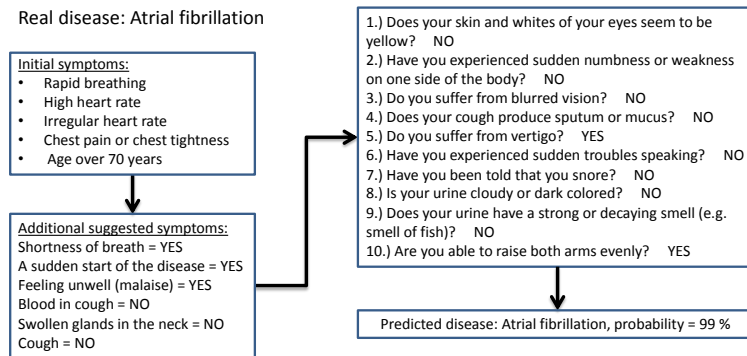


Fig. 2. Example of testing the diagnostic method on a real patient with atrial fibrillation.

4 Discussion and Conclusion

The initial results, based on both training and testing the method on virtual patients, show very high sensitivity, specificity, and classification accuracy (all over 80%). These values are probably too optimistic, according to the opinion of our medical associates. One of the main reasons for these results is that both the training and the testing set were generated from the same expert table. Because virtual patients are not biased when answering the questions, it is even more necessary to train and test the method on real patients, which we plan to do in the future. Moreover, the medical conditions were classified only into 15 different classes, which is far below the number of possible medical conditions in reality. The predefined medical conditions are also very distinctive in terms of symptom manifestation and it is therefore easier to distinguish between them (higher classification accuracy). The exceptions here are chronic obstructive pulmonary disorder, pneumonia, tuberculosis, and sleep apnoea; they are more frequently misclassified due to the similarity of their symptoms. In the future, we intend to incorporate hierarchical classification (e.g. additional class 'pulmonary disease'), when the data is insufficient for reliable differentiation between similar medical conditions. Current implementation of the method utilizes only the questionnaire for all of the symptom. In the future, we will use actual sensor input to determine the presence of a few specific symptoms. Additionally, we plan to include a larger number of medical conditions and implement intelligent methods for multilabel classification for discovering combinations of conditions.

Acknowledgments. We would like to thank MESI, DLabs, Gigodesign, Faculty of Medicine, and other partners, collaborating in the *MESI Simplifying diagnostics* team for the Qualcomm Tricorder XPRIZE competition.

References

1. Qualcomm Tricorder XPRIZE, <http://www.qualcommtricorderxprize.org/>
2. MESI, <http://www.mesimedical.com/home/>
3. McCormick T.H., Rudin C. and Madigan D.: A Hierarchical Model for Association Rule Mining of Sequential Events: an Approach to Automated Medical Symptom Prediction. *Annals of Applied Statistics*. (2012)
4. Soni S. and Vyas O.P.: Fuzzy Weighted Associative Classifier: A Predictive Technique For Health Care DataMining. *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)*, Vol. 2, No.1, (2012)
5. Soni S. and Vyas O.P.: Using Associative Classifiers for Predictive Analysis in Health Care Data Mining. *International Journal of Computer Applications*, Vol. 4, No. 5, (2010)
6. Peng H., Ding C.: Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *Journal of Bioinformatics and Computational Biology*, Vol. 3, No. 2, (2005)
7. Somrak M., Luštrek M., Šušterič J., Krivic T., Mlinar A., Travnik T., Stepan L., Mavsar M., Gams M.: Tricorder: Consumer Medical Device for Discovering Common Medical Conditions. *Informatica* 38, Ljubljana (2014)