MINING TELEMONITORING DATA FROM CONGESTIVE-HEART-FAILURE PATIENTS

Mitja Luštrek[1,2], *Maja Somrak*[1,2]
[1]Jožef Stefan Institute, Department of Intelligent Systems
[2]Jožef Stefan International Postgraduate School
e-mail: {mitja.lustrek, maja.somrak}@ijs.si

ABSTRACT

The Chiron project carried out an observational study in which congestive-heart-failure patients were telemonitored in two countries. Data from 1,068 recording days of 25 patients were gathered, consisting of 15 dynamic parameters (measured daily or continuously) and 49 static parameters (measured once or a few times during the study). The features derived from these parameters were mined for their association with the feeling of good/bad health. The findings mostly correspond to the current medical knowledge, although some may represent new insights.

1 INTRODUCTION

Telemonitoring of patients with chronic diseases is becoming technically increasingly feasible, but benefits for the patients are not always apparent, nor is it clear how to make the most of the data obtained this way. In the case of heart failure, two systematic literature reviews showed lower mortality resulting from telemonitoring [1][2], but in the trials they reviewed, telemonitoring was mostly compared with conventional care worse than what is offered today. Conversely, two large recent trials showed no benefit from telemonitoring [3]. However, the telemonitoring in these two trials was not very advanced – the monitored parameters were limited and no intelligent computer analysis was involved. We can conclude from this that as the conventional care improved, so should telemonitoring. One way to do so is by using intelligent computer methods on the gathered data, both to save the time of the medical personnel who would otherwise have to look at all the data themselves, and to uncover previously unknown relations in the data.

This paper describes the mining of telemonitoring data from congestive-heart-failure (CHF) patients gathered in the Chiron project [4]. The objective of this project was to develop a framework for personalized health management with a focus on telemonitoring. The Chiron patients were equipped with a wearable ECG, activity, body-temperature, sweating and sensors. In addition, their blood pressure, blood oxygen saturation, weight, and ambient temperature and humidity were measured [5]. The data gathered this way was fed into a decision-support system, whose objective was to estimate the health risk of the patients [6]. However, since there is not enough knowledge on how to associate the values of the various measured parameters with the risk, an observational study was carried out in the project with the intention to generate such knowledge. This paper presents an initial analysis of the data gathered in this study.

2 DATA FROM THE CHIRON STUDY

2.1 Data gathering and description

The data analyzed in this paper were gathered in the period from May 2013 to May 2014. The whole study included 38 CHF patients: 19 from the United Kingdom and 19 from Italy. However, some of the data were incomplete, so only the data of 12 patients from the UK and 13 patients from Italy were included in the analysis. These 25 patients together provided a total of 1,068 usable recording days. The data consists of 64 parameters carefully selected based on their relevance to CHF [7].

The initial measurements of 49 *static* parameters were taken for each of the patients at the beginning of the study. This data includes general patient information (age, gender, BMI, waist-to-hip ratio, smoking, etc.), their current medical treatments (beta blockers, anti-coagulants, ACE inhibitors, etc.), related health conditions (arrhythmias, hypertension, diabetes, etc.) and the results of a blood analysis (hemoglobin, lymphocytes, LDL/HDL cholesterol, blood glucose, Na and K levels, etc.). Some of these measurements were repeated periodically every few weeks to provide up-to-date information. However, the exact period varied from patient to patient and roughly half of the patients only had the measurements taken at the beginning of the study.

During the study, the patients were wearing vital-signs monitoring equipment [5] for several hours each day. The equipment consisted of an ECG device, two accelerometers places on the chest and thigh, a body-temperature and a humidity sensor. The ECG recordings were subsequently analyzed to extract the physiological parameters related to the heart rhythm: heart rate, QRS interval, QT interval, PR interval, T wave amplitude and R wave amplitude. The accelerometers continuously provided the patient's activity and energy-expenditure estimation. The temperature and humidity sensors provided the measurements of the skin temperature and sweating index in five-minute intervals.

The patients were also provided with a mobile application for generating weekly and daily reports. The patients reported their overall feeling of health with respect to the previous day on a daily basis (feeling much worse than yesterday, worse, the same, better or much better), and answered 13 questions about their health and well-being on a weekly basis. In addition, they reported measurements of systolic and diastolic blood pressure, body mass, blood oxygen saturation, and ambient temperature and humidity. These – together with the continuously monitored parameters – are labeled *dynamic* in Section 3.

The study also intended to gather data about hospital admissions and deaths, but no such events occurred during the study period. Therefore we decided to use the patients' self-reports of health instead. The analysis in this paper is based on the daily questions about the feeling of health.

2.2 Data preprocessing

The ECG and accelerometer data recordings required the most attention when preprocessing the data prior to the data mining. These two types of recordings also generated the vast majority of all the gathered data.

The ECG signal was already processed with the Falcon algorithm [5], producing an output where each heart beat is described with an 11-tuple. Because the tuples were not explicitly separated and some of them are incomplete, it was important to distinguish between them in order to extract the specified parameters. We used R-peaks in the ECG signal to identify distinct tuples. Additionally, a lot of the data was corrupt or missing, so those parts had to be removed.

Similar problems occurred when processing the accelerometer data. It was not possible to extract the information about the activity and energy expenditure if a recording of any one of the axes of either of the two sensors was missing. If a patient forgot to wear both sensors, or one had an empty battery, the data thus had to be discarded.

Finally, some data was not uploaded successfully to the servers due do connection problems, and some data are missing as a result of inconsistent patients' behavior.

All of the parameters that were measured continuously were further separated by the main activities of the day: during lying, sitting and moving separately (resulting in features labeled *per_act* in Section 3) or during all the activities together (*all_act*). The ratios of the durations of these three activities were calculated for each day. For every parameter that was measured continuously or multiple times per day, the average value (*avg*) and standard deviation (*sd*) were calculated; the calculations were done for separate activities and for the whole day.

The key value whose association with the other monitored parameters we study in this paper – the overall feeling of health – was reported by the patients relatively to the previous day. Since the value is not absolute (e.g., feeling well) but relative (e.g., feeling better or worse than yesterday), it is associated with the measurements of both the current and the previous day. Because of that we introduced features that represent changes of the parameters' values with respect to the previous day (*chg*). Again, the calculations were done for separate activities and for the whole day.

For the purpose of data mining, classes were assigned to the data. If each of the five distinct feelings of health corresponds to one class, the differences between them are too small. Therefore we decided to have only two classes:

- Much worse vs. much better (*MW-MB*)
- Much worse or worse three times in a row vs. much better or better three times in a row (*MW3-MB3*)
- Much worse or worse vs. much better or better (*MWW-MBB*)
- Much worse vs. everything else (*MW-E*)
- Much worse or worse three times in a row vs. everything else (*MW3-E*)
- Much worse or worse vs. everything else (*MWW-E*)

The majority of the data instances have the class 'feeling the same as yesterday', while very few instances have 'feeling much better' or 'feeling much worse'. Because of this, the first three classes result in discarding the majority of the instances (only 69, 101 or 285 instances remain), while the last three use all 1,086 of them. Since classes are imbalanced, particularly in the last three cases, we used cost-sensitive classification, with the costs of misclassifications compensating for the imbalances.

3 MINING THE DATA

Since the number of combinations of data-mining algorithms, features and classes is huge, we designed a three-step data-mining procedure (described in detail in Sections 3.1–3.3):

1. Selection of algorithms that classify the data with a high accuracy and yield understandable models
2. Using the selected algorithms, selection of features that classify the data with a high accuracy and are understandable
3. Using the selected algorithms and features, selection of classes that result in accurate models

At the end of these three steps, we ended up with a number of interesting models, some of which are presented in Section 3.4.

3.1 Selection of algorithms

In the first step we used MW3-MB3 classes and the avg subset of dynamic all_act features. We compared several algorithms from the Weka suite [8] shown in Table 1. We selected the underlined algorithms for the experiments in Sections 3.2 and 3.3 due to their accuracy and in the case of JRip to have another understandable algorithm.

Table 1: Comparison of data-mining algorithms

Algorithm	Accuracy
Random Forest	79.3 %
Naive Bayes	77.4 %
J48	76.3 %
SVM, Puk kernel	74.5 %
SVM, linear kernel	74.2 %
SGD	73.8 %
Multilayer Perceptron	73.2 %
JRip	71.9 %
kNN, k = 1	60.9 %
kNN, k = 2	56.2 %
kNN, k = 3	47.8 %
SVM, RBF kernel	40.1 %

3.2 Selection of features

We first compared predefined features sets described in Section 2. Since the number of combinations is large, we proceeded in several sub-steps. First, we compared subsets of dynamic all_act features, finding that only avg and avg + chg subsets performed better than the rest. The results are shown in the first segment of Table 2 with the highest accuracy for each algorithm in bold. Second, we added per_act features to these two subsets of features, finding the extended features worse than all_act features alone (second segment of Table 2). And third, we combined these two subsets of features with static features, finding them best of all (third segment of Table 2). However, given the small number of patients, it is likely that the static features identified individual patients instead of taking into account their general characteristics. Because of that we retained all the underlined features for experiments in Section 3.3.

Table 2: Comparison of predefined feature sets

Features \ Algorithm	Naive Bayes	SVM, Puk	JRip	J48	Random Forest
Dynamic, all_act, avg + chg	75.5	**80.0**	70.6	**76.9**	**80.3**
Dynamic, all_act, avg	**77.4**	74.5	**71.9**	76.3	79.3
Dynamic, all_act, avg + sd	75.3	73.1	70.9	73.3	77.7
Dynamic, all_act, avg + chg + sd	74.0	78.7	70.3	75.2	78.3
Dynamic, all_act, chg + sd	67.1	78.6	64.6	64.9	71.9
Dynamic, all_act, chg	62.1	71.2	55.5	64.8	64.4
Dynamic, all_act, sd	58.2	65.4	63.0	64.6	66.9
Dynamic, all_act + per_act, avg	**77.0**	**72.5**	**71.6**	75.7	78.4
Dynamic, all_act + per_act, avg + chg	73.4	71.8	71.0	**76.7**	**79.1**
Dynamic + static, all_act, avg	77.5	79.2	75.5	76.4	79.3
Dynamic + static, all_act, avg + chg	**77.8**	**80.4**	**77.0**	**79.6**	**80.5**

We also tested automatic feature selection methods from the Weka suite. None of the methods performed well on its own, so we used the features selected by at least two methods out of the following: Correlation-based Feature Subset, Gain Ratio, ReliefF, Symmetrical Uncertainty and Wrapper (the end result of the Wrapper approach was the union of features selected when each of the five algorithms selected in Section 3.1 were used). As the starting point, we used all features, all dynamic features, and avg + chg subset of all_act dynamic features. The results in Table 3 show that the first and third of these starting points resulted in the best models obtained so far, although we retained all the underlined features for the experiments in Section 3.3.

Table 3: Comparison of automatic feature selection

Features \ Algorithm	Naive Bayes	SVM, Puk	JRip	J48	Random Forest
All features, FS	75.5	**80.0**	70.6	**76.9**	**80.3**
Dynamic, all_act, avg + chg, FS	**77.4**	74.5	**71.9**	76.3	79.3
Dynamic + static, all_act, avg + chg	75.3	73.1	70.9	73.3	77.7
Dynamic, all_act, avg + chg	74.0	78.7	70.3	75.2	78.3
Dynamic, all_act, avg	67.1	78.6	64.6	64.9	71.9
Dynamic, FS	62.1	71.2	55.5	64.8	64.4

3.3 Selection of classes

We compared the accuracies of different classes on all the algorithms selected in Section 3.1 and all the features selected in Section 3.2. In Table 4 we report the F-measure for the Random Forest algorithm (most accurate overall), averaged over all the features. The F-measure was chosen because of the class imbalance, particularly for the three 'vs. everything else' pairs of classes. One can see that MW3-MB3 performed best, probably because it strikes the best balance between the difference between the two classes in the pair, and the number of instances in the dataset. MW-MB may have too few features, while in the other cases the difference between the two classes is too small.

Table 4: Comparison of classes

Classes	MW-MB	MW3-MB3	MWW-MBB	MW-E	MW3-E	MWW-E
F-measure	0.77	**0.79**	0.66	0.55	0.56	0.61
Instances	69	101	285	1,068	1,068	1,068

3.4 Interesting models

Classification models were built with the J48 and JRip algorithms (being the most understandable of the five selected in Section 3.1) on all the features selected in Section 3.2. Two examples are presented in Figure 1 and Figure 2. They show that a high heart rate (*HR_avg_all_activities* in the figures) and short QRS interval (*QRS_avg_all_activities*, a feature of the ECG signal) are associated with the feeling of good health, which corresponds to the existing medical knowledge. Increased weight (*DRWChg*) is associated with bad health, which makes sense, since it often signifies excess fluid retention, a common problem of CHF patients. Low humidity (*HumA*) and decrease in humidity (*HumAChg*) are associated with good health, which matches the medical opinion that CHF patients often badly tolerate humid weather, although there is little hard evidence for this. Oxygen saturation (*DRS*) below 97 % is associated with bad health in the second model, which is normal, since the saturation in healthy individuals is 96 % – 100 %. Finally, the first model associates high systolic blood pressure (*SBP*) and the second low diastolic blood pressure (*DBP*) with good health. This is expected in CHF patients, since their hearts have problems pumping out enough blood (low systolic blood pressure) as well accepting enough blood (high diastolic blood pressure).

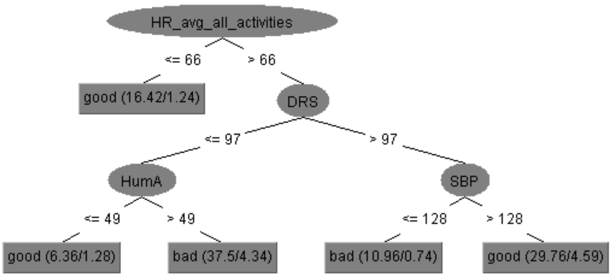

Figure 1: J48 classification tree on the avg subset of all_act dynamic features

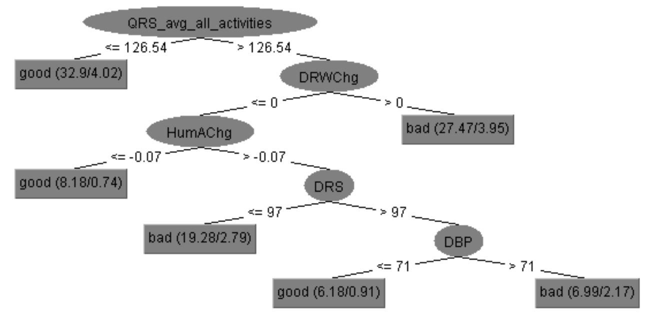

Figure 2: J48 classification tree on the avg + chg subset of all_act dynamic features

4 CONCLUSION

Telemonitoring can provide huge quantities of medically relevant data, which has the potential to revolutionize the care of patients with chronic diseases. However, before this can happen, the data must be properly interpreted, for which the current knowledge is not yet entirely adequate. This paper presents the data gathered by telemonitoring of CHF patients, and the first attempt to uncover interesting relations in the data by data mining. A systematic procedure for the selection of appropriate data-mining algorithms, features and classes was designed, whose output were a number of models associating telemonitored parameters with the feeling of good or bad health. The models correspond quite well to the current medical knowledge, which demonstrates the validity of our approach.

In the future, we need to solve the technical difficulties with extracting the ECG parameters and compute some new features that may be relevant (e.g., QT interval prolongation, a feature of the ECG signal that is known to be associated with cardiovascular problems). Furthermore, the models resulting from data mining must be carefully examined by cardiologists, both the models presented in the paper and others. Those that contain hitherto unknown relations may be even more important than those that correspond to the current medical knowledge, since the relations in them may yield new and important insights. Finally, it would be desirable to study data that contain events such as hospital admissions or even deaths, since the findings on such data would be more reliable than on data that only contains self-reported feeling of health. However, another observational study would be needed for that, which is a difficult proposition that would require substantial funding.

Acknowledgement

This research described in this paper was carried out in the Chiron project, which was co-funded by the ARTEMIS Joint Undertaking (grant agreement # 2009-1-100228) and by national authorities.

References

[1] C. Klersy, A. De Silvestri, G. Gabutti, F. Regoli, A. Auricchio. A meta-analysis of remote monitoring of heart failure patients. *Journal of the American College of Cardiology* 54, 2009, pp. 1683–1694.

[2] S. C. Inglis, R. A. Clark, F. A. McAlister, S. Stewart, J. G. Cleland. Which components of heart failure programmes are effective? A systematic review and meta-analysis of the outcomes of structured telephone support or telemonitoring as the primary component of chronic heart failure management in 8323 patients: abridged Cochrane Review. *European Journal of Heart Failure* 13, 2011, pp. 1028–1040.

[3] C. Sousa, S. Leite, R. Lagido, L. Ferreira, J. Silva-Cardoso, M. J. Maciel. Telemonitoring in heart failure: A state-of-the-art review. *Revista Portuguesa de Cardiologia* 33 (4), pp. 229–239.

[4] Chiron project. http://www.chiron-project.eu/.

[5] E. Mazomenos, J. M. Rodríguez, C. Cavero, G. Tartarisco, G. Pioggia, B. Cvetković, S. Kozina, H Gjoreski, M. Luštrek, H. Solar, D. Marinčič, J. Lampe, S. Bonfiglio, K. Maharatna. Case Studies. In *System Design for Remote Healthcare*, 2014, pp. 277–332.

[6] M. Luštrek, B. Cvetković, M. Bordone, E. Soudah, C. Cavero, J. M. Rodríguez, A. Moreno, A. Brasaola, P. E. Puddu. Supporting clinical professionals in decision-making for patients with chronic diseases. *Proc. IS 2013*, pp. 126–129.

[7] P. E. Puddu, J. M. Morgan, C. Torromeo, N. Curzen, M. Schiariti, S. Bonfiglio. A clinical observational study in the Chiron project: Rationale and expected results. In *Impact Analysis of Solutions for Chronic Disease Prevention and Management*, 2012, pp. 74–82.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations* 11 (1), 2009, pp. 10–18.