# Predicting Health Patterns Using Sensor Sequence Similarity and NLP

Zahra Hajihashemi
Computer Science Department
University of Missouri-Columbia
Columbia, Missouri, USA
Zhmr5@mail.missouri.edu

Mihail Popescu
Health Management and Informatics Department
University of Missouri-Columbia
Columbia, Missouri, USA
popescum@health.missouri.edu

*Abstract*— **Health information technology has been used in long-term care to improve outcomes and reduce cost. In Tiger Pace, an aging in place facility from Columbia, MO, we deployed sensor networks together with an electronic health record (EHR) system to provide early illness recognition. In this paper, we describe a methodology for early illness based on non-wearable sensor data and concepts extracted from nursing notes using Natural Language Processing (NLP). The methodology is inspired from genomic sequence annotation using BLAST. First, we extract a set of Unified Medical Language System (UMLS) concepts from each nursing note using Metamap, a NLP tool provided by UMLS. Then, we associate each daily sensor sequence with the medical concepts related to the nursing notes issued that day for that patient. Finally, to infer the health concepts for an unknown day, we compute the similarity between its sensor sequence and those available in the database. The challenges presented by this method are finding the most suitable multi-attribute time sequence similarity and aggregation of the retrieved concepts. On a pilot dataset from three Tiger Place residents, with a total of 1685 sensor days and 358 nursing records, we obtained an average precision of 0.34 and a recall of 0.52.**

*Index Terms*— **Eldercare monitoring, early illness recognition, natural language processing (NLP), health context aware algorithms.**

## I. Introduction

Eldercare has become more challenging as older adults prefer to live independently for as long as they are able regardless of conditions such as frailty, dementia and risk of falling. Late health assessment is an aggravating risk factor that usually occurs because of fear of being institutionalized and the failure of physician's assessments [1].

Sensor networks have been used in last decades as a promising solution to monitor older adult health [2, 3]. MIT's PlaceLab, Georgia Tech's Aware House and Honeywell's Independent Lifestyle Assistant are successful examples [4, 5, 6, 7]. Some health monitoring systems can detect and estimate activity patterns and assess medication compliance using a collection of sensors and predictive algorithms [8, 9]. Considering the health context of the monitored patient is still an unsolved problem.

In this paper we have leveraged the work performed in Tiger Place, our living laboratory, where sensor networks have been installed in the home of residents since the fall of 2005 [10]. Here, we employ sensor data and contextual health information such as chronic conditions and nursing comments provided by EHR to identify health patterns. In previous work [11] we used sensor data to identify early signs of illness and send alerts to clinical staffs that provide feedback on the clinical relevance of each alert.

In this paper, we describe a methodology for predicting early signs of illness based on sensor data similarity and UMLS concepts extracted from related nursing notes. This paper is organized as follows. In Section II we describe the system architecture and available sensors data. Section III presents our method to predict medical concepts based on sensor data similarity. Section IV shows experiments and results. Finally, in section V we give conclusions and future work.

## II. System Architecture

We deployed our integrated monitoring system (see figure 1) in 36 Tiger Place apartments. All deployed sensors are non-wearable. The monitoring has been ongoing since fall 2005 with an average monitoring time of nearly 2 years per resident.

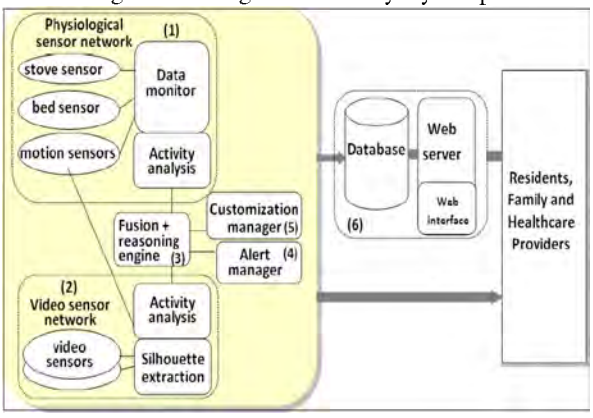


Fig.1. Tiger Place Sensor Network Architecture.

Various sensors have been deployed in each apartment: motion, radar, Microsoft kinect and bed. The bed sensor is able to measure bed motion, pulse and breathing. Each sensor sends an X10 signal that is logged together with a time stamp in our

sensor database. In this study we use only motion, pulse, breathing and restlessness sensor data. Table 1 shows the sensor data available for this study from three Tiger Place residents. The data was aggregated hourly for each day. For each patient we also collected visit notes that describe various physical and/or emotional issues inputted by the nursing personnel in the EHR. Note that there fewer notes than sensor data, as some days didn't have any nursing comment

TABLE I. TIGER PLACE DATASET

| Resident Code | Number of sensor days | Number of comment days |
|---|---|---|
| 1 | 440 | 83 |
| 2 | 745 | 44 |
| 3 | 500 | 499 |

## III. METHOD

Many sequence similarity algorithms have been proposed in the literature [12,13,14,15,16]. One approach is to use the Euclidean distance to measure the similarity between two sequences. In this approach, a sequence is considered as a point in an appropriate multi-dimensional space. Non-Euclidian metrics have been used as another approach for computing the similarity of time sequences [17,18]. Haar wavelet transform have been used to reduce the dimensionality of the time sequences [19]. The shift and scale comparisons of the sequences have been used in another study to allow comparisons under different experimental conditions [20].

In our study, health prediction is based on the similarity of the sequence data provided by four sensors: motion, restlessness, pulse and breathing. However, considering each sensor data separately would not be helpful. As a proof, figure 2 shows the results of iVAT algorithm on the sensor data set. The iVAT algorithm has been used to visualize the different possible clusters in the data [21]. The number of different color tones demonstrates the number of possible clusters. As figure 2 shows, there are no meaningful clusters in any dimension if we consider each sensor separately. Instead, here we used a simple similarity measure based on root mean square (RMS) which is widely used in signal processing. For two sensor sequences $X = \{x_{ij}\}$, $Y = \{y_{ij}\} \in R^4 \times R^{24}$ we can compute a distance $d(X,Y)$ as:

$$d(X,Y) = \sqrt{\frac{1}{24}\left((X_1 - Y_1)^2 + \cdots + (X_{24} - Y_{24})^2\right)}, \quad (1)$$

where

$$X_i = \sqrt{\frac{1}{4}\left(x_{1i}^2 + x_{2i}^2 + x_{3i}^2 + x_{4i}^2\right)} \quad (2)$$

$$Y_i = \sqrt{\frac{1}{4}\left(y_{1i}^2 + y_{2i}^2 + y_{3i}^2 + y_{4i}^2\right)}$$

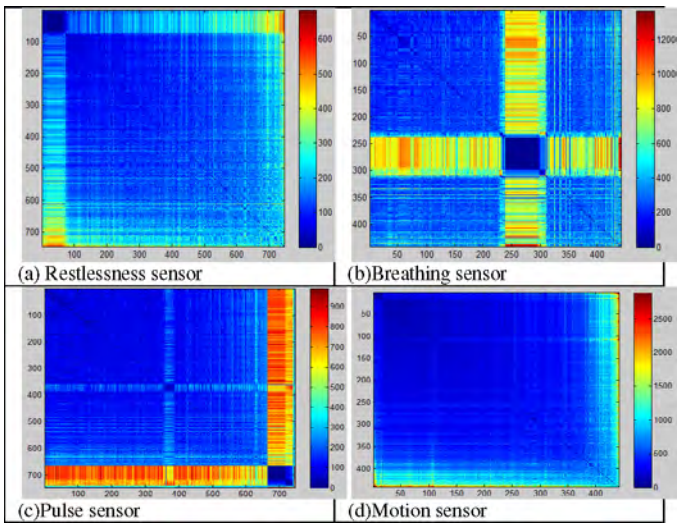


Fig.2. Results of iVAT algorithms for patient 3003.

Our main goal is to predict patient's health status using medical concepts from similar sequences. To extract the medical concepts from the nursing visit notes we parsed the free text data using the Metamap NLP tool provided by UMLS (http://metamap.nlm.nih.gov/). Metamap associates each medical concept found in the nursing notes to a Concept Unique Identifier (CUI) from UMLS.

To predict possible health problems given a sensor sequence $Z \in R^4 \times R^{24}$, we select the CUI's associated to the sequence $X_{best}$ most similar to Z, i.e $X_{best} = \min_n\{d(Z,X)\}$. We evaluated our results in terms of precision/recall, where precision is the fraction of retrieved concepts that are relevant, while recall is the fraction of relevant concepts that are retrieved. We only include in our experiments the days that have comments attached (column 3 in Table I). Alternatively, we could label the days with no comments in the EHR as "normal" and include them in the experiment. However, we left this idea for future work. We performed the experiments using a leave-one-one scheme and averaged the precision and recall for the entire resident dataset.

## IV. EXPERIMENTAL RESULTS

Figure 3 shows an example of two similar sequences for patient #2. As we can see in this figure, analyzing each sensor data separately might not result in a good match.

Table II demonstrates the results of our illness prediction experiments on the pilot dataset from Table I.

TABLE II. RESULTS OF THE ILLNESS PREDICTION EXPERIMENTS

| Resident Code | Precision | Recall | F-measure | Number of Days |
|---|---|---|---|---|
| #3 | 0.4256 | 0.7446 | 0.5388 | 283 |
| #2 | 0.2897 | 0.3559 | 0.3174 | 22 |
| #1 | 0.3361 | 0.4639 | 0.3825 | 53 |

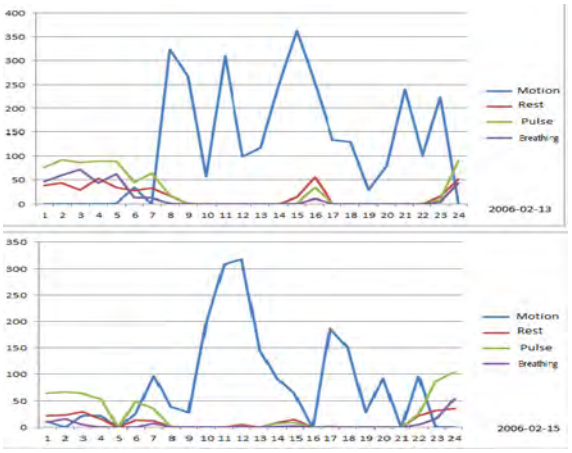


Fig.3. Two similar sequences for patient #2.

From Table II we see that the precision and recall are reasonable when plenty of annotated sequences are available (as is the case with resident #3). While the lack of comments might be a problem for relatively healthy nursing home residents (such as #1 and #2), we believe that in general it doesn't apply for the typical nursing home resident that has multiple chronic diseases (hence many encounters with the nursing personnel).

## V. CONCLUSION

We have shown potential of identifying early changes in health using in-home sensors and NLP of the nursing notes. We believe our method can scale well to multiple residents and multiple nursing homes. This research leverages ongoing work at the University of Missouri-Columbia (MU) in the use of sensor technology for in-home health assessment. We integrated our sensor networks with a home-grown nursing EHR and investigated health context aware computational algorithms for health assessments.

While our work showed promising results, we acknowledge that we left many questions unanswered. First, is it possible to extend sequence similarity across residents? What would be the criteria for doing so? Here, we used a very simple multi-attribute sequence distance. What would be the best multi-attribute sequence similarities (distances) for our problem? Also, we used only the best matching sequence to infer the unknown medical concepts. What if multiple sequences (from multiple days) are used for inference? How can we aggregate them?

## ACKNOWLEDGMENT

This project has been funded by a NSF SHWB grant, award IIS-1115956.

## REFERENCES

[1] TL Hayes, M Pavel, JA Kaye, "An unobtrusive in-home monitoring system for detection of key motor changes preceding cognitive decline", Proc. of the 26th Annual Intl. Conf. of the IEEE EMBS, San Francisco, CA, 2004, pp. 2480-2483.

[2] J Rowan, ED Mynatt, "Digital Family Portrait field trial: support for Aging in Place", Proc. of the SIGCHI Conf. on Human Factors in Computing Systems, New York,2005, pp. 521-530, ACM Press.

[3] P. Cuddihy, J. Weisenberg, C. Graichen, M. Ganesh, "Algorithm to automatically detect abnormally long periods of inactivity in a home", Proc. of the 1st ACM SIGMOBILE Intl. Workshop, New York, 2007, pp. 89-94.

[4] S. Intille, K. Larson, E. Munguia-Tapia, J. Beaudin, P. Kaushik, J. Nawyn, et al, "Using a live-in laboratory for ubiquitous computing research", Berlin: Proc. of Pervasive, 2006.

[5] CD. Kidd, RJ. Orr, GD. Abowd, IA. Atkeson, B. Essa, B. MacIntyre, et al., "The Aware Home: A living laboratory for ubiquitous computing research", Proc. of the 2nd International Workshop on Cooperative Buildings-CoBuild'9, 1999.

[6] KZ. Haigh, LM. Kiff, G. Ho, "Independent Lifestyle Assistant: Lessons learned", Assistive Technology, 18, 2006, pp.87-106.

[7] D. Heise, M. Skubic, "Monitoring pulse and respiration with a non-invasive hydraulic bed sensor", Proc.,32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Buenos Aires, Argentina, 2010.

[8] R. Beckwith, "Designing for ubiquity: The perception of privacy", Pervasive Computing, 2003, pp.40-6.

[9] D. Mack, M. Alwan, B. Turner, R. Suratt, R. Felder, "A passive and portable system for monitoring heart rate and detecting sleep apnea and arousals: Preliminary validation. Arlington VA", Proceedings Transdisciplinary Conference on Distributed Diagnosis and Home Healthcare (D2H2), 2006.

[10] MJ. Rantz, K. Dorman-Marek, M. Aud, HW. Tyrer, M. Skubic, G. Demiris, et al., "A technology and nursing collaboration to help older adults age in place", Nursing Outlook; 2005, pp.40-45.

[11] M. Skubic, G. Alexander, M. Popescu, M. Rantz, J. Keller, "A Smart Home Application to Eldercare: Current Status and Lessons Learned", Technology and Health Care, 17, 2009, pp. 183-201.

[12] R. Agrawal, C. Faloutsos, and A. Swami; "Efficient similarity search in sequence databases"; In FODO, Evanston, Illinois, October 1993.

[13] K.-P. Chan and A.W.-C. Fu; "Efficient time series matching by wavelets"; In ICDE, 1999.

[14] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos; " Fast subsequence matching in time-series databases"; In SIGMOD, pp. 419–429,

[15] T. Kahveci and A. Singh; "Variable length queries for time series data"; In ICDE, Heidelberg, Germany, 2001.

[16] C. Shahabi, X. Tian, and W. Zhao; "TSA-tree: A wavelet-based approach to improve the efficiency of multi-level surprise and trend queries"; In SSDBM, 2000.

[17] R. Agrawal, K. Lin, H.S. Sawhney, and K. Shim;"Fast similarity search in the presence of noise, scaling, and translation in time-series Databases"; In VLDB, Z˙urich, Switzerland, September 1995.

[18] S.-L. Lee, S.-J. Chun, D.-H. Kim, J.-H. Lee, and C.-W. Chung; "Similarity search for multidimensional data sequences"; In ICDE, San Diego, CA, 2000.

[19] D. Rafiei and A.O. Mendelzon; " Efficient retrieval of similar time sequences using DFT"; In FODO, Kobe, Japan, 1998.

[20] Tamer Kahveci, Ambuj Singh and Aliekber Gürel; "Similarity searching for multi-attribute sequences"; In proc. SSDBM. 2002, pp. 175-184.

[21] T.C. Havens and J.C. Bezdek (2012). An efficient formulation of the improved visual assessment of tendency (iVAT) algorithm. IEEE Trans. Knowledge and Data Engineering, 2005, pp.813-822.