Early Illness Recognition Using In-home Monitoring Sensors and Multiple Instance Learning

M. Popescu[1]; A. Mahnot[2]
[1]Health Management and Informatics, University of Missouri, Columbia, MO, USA;
[2]Electrical and Computer Engineering, University of Missouri, Columbia, MO, USA

Keywords
Automated pattern recognition, multiple instance learning, eldercare

Summary
Background: Many older adults in the US prefer to live independently for as long as they are able, despite the onset of conditions such as frailty and dementia. Solutions are needed to enable independent living, while enhancing safety and peace of mind for their families. Elderly patients are particularly at-risk for late assessment of cognitive changes.
Objectives: We predict early signs of illness in older adults by using the data generated by a continuous, unobtrusive nursing home monitoring system.
Methods: We describe the possibility of employing a multiple instance learning (MIL) framework for early illness detection. The MIL framework is suitable for training classifiers when the available data presents temporal or location uncertainties.
Results: We provide experiments on three datasets that prove the utility of the MIL framework. We first tuned our algorithms on a set of 200 normal/abnormal behavior patterns produced by a dedicated simulator. We then conducted two retrospective studies on residents from the Tiger Place aging in place facility, aged over 70, which have been monitored with motion and bed sensors for over two years. The presence or absence of the illness was manually assessed based on the nursing visit reports.
Conclusions: The use of simulated sensor data proved to be very useful for algorithm development and testing. The results obtained using MIL for six Tiger Place residents, an average area under the receiver operator characteristic curve (AROC) of 0.7, are promising. However, more sophisticated MIL classifiers are needed to improve the performance.

Correspondence to:
Mihail Popescu, PhD
Health Management and Informatics
University of Missouri
HMI Department
324 Clark Hall
Columbia, MO 65211
USA
E-mail: popescum@missouri.edu

Methods Inf Med 2012; 51: 359–367
doi: 10.3414/ME11-02-0042
received: November 7, 2011
accepted: July 9, 2012
prepublished: July 20, 2012

1. Introduction

Many older adults in the US prefer to live independently for as long as they are able, despite the onset of conditions such as frailty and dementia. Solutions are needed to enable independent living while enhancing safety and peace of mind for their families [3, 15].

Aging adults may sometimes purposefully mask any decline in abilities to avoid outside intervention or concern held by their children [15]. Elderly patients are particularly at-risk for late assessment of cognitive changes due to many factors: their impression that such changes are simply a normal part of aging, their reluctance to admit to a problem, their fear of being institutionalized and even the failure of physicians to fully assess their cognitive function due to the belief that no intervention is possible [7].

The above observations suggest the need for automatically detecting early signs of illness and alerting the health care provider in a timely manner [10]. It has been shown that diseases such as cardiac arrhythmia, congestive heart failure and pneumonia, among others, may produce a sudden onset of anxiety [2]. Signs of anxiety such as restlessness, insomnia, frequent urination or diarrhea [5] translate into observable behavior changes such as abnormal sleep or room motion patterns. Our early illness approach is based on the assumption that the abnormal behavior patterns can be captured by the environmental sensors (e.g. movement and bed sensors) that we currently have deployed in Tiger Place [16].

We note that the algorithms we develop in this paper attempt to model behavior rather than physiology. While physiology is very similar among humans, the behavior is not. This implies that, while we can train classifiers with data from a large amount of different patients, the same is not true for behavioral data. Only sensor data from a given patient can be used to predict his/her behavior. As a result, in behavior prediction experiments, the amount of temporal data is more important than the number of available patients (sample size).

Our sensor data capture external (behavioral) information about the resident that is then linked to existent medical records or self reported health status. In previous work [11] we predicted elevated pulse pressure (pulse pressure = systolic pressure – diastolic pressure) based on sensor data. For that application we had suitable data such that two-class classifiers (e.g. SVM and neural net)

could be trained. In subsequent work [12] we tried to predict "abnormal events" based on sensor data. In that work, due to the strong class imbalance present in the data (not many abnormal event labels available), we decided to employ a one-class classifier approach. In this paper, an extended version of the work presented at the IDAMAP 2010 symposium [13], we show how a MIL framework can be employed in conjunction with data provided by unobtrusive sensors deployed in the living environment to detect early signs of illness based on health status extracted from nursing visit reports.

Multiple instance learning (MIL) [4, 8, 9] is a supervised learning approach which can be applied in a setting in which individual labels for each training example are either hard to assign (e.g. labeling objects of interest in an image) or not available (e.g. in which hour of the day the resident did not feel well). Instead, it is much easier to obtain labels for sets of objects (called bags), e.g. labeling the whole image as a "positive" example or labeling the whole day as "bad". MIL has been successfully employed in applications such as scene recognition [9], image retrieval [18] and drug-target interaction [4]. A brief introduction to MIL is given in the next section.

2. Methods

2.1 Multiple Instance Learning

In multiple instance learning, classifiers are trained with labeled sets of instances called "bags". Each positive bag, B_i^+, contains at least one positive instance. Theindividual labels of the instances in each bag are not known at training time. A negative bag, B_i^-, contains (theoretically) only negative instances. In our case, a bag consists of 24 vectors of sensor data that correspond to the 24 hours before the nurse report. Due to the lack of information, we considered that each nursing visit happened at 12 pm. The days in which the nurse report revealed a health event were labeled "positive" (i.e. they contain some abnormal behavior). The days in which the nurse report did not mention any health problems were considered "negative".

There are many MIL algorithms. In this paper we used the diversity density (DD) framework proposed in [8]. The DD of a point x in feature space, $x \in R^p$, is proportional to the number of positive bags with instances close to x and to the number of negative bags with instances far from x.

If we denote $B_{ij}^+ \in R^p$ the j-th instance of the i-th positive bag and B_{ij}^- the j-th instance of the i-th negative bag we can find the point x_{opt} that maximizes DD as:

$$\underset{x}{\operatorname{argmax}} \prod_i P(x \mid B_i^+) \prod_i P(x \mid B_i^-), \quad (1)$$

where, for example, $P(x|B_i^+)$ is computed as:

$$P(x|B_i^+) = 1 - \prod_j \left[(1 - P(x|B_{ij}^+)\right] \quad (2)$$

where $P(x|B_{ij}^+)$ can be computed using a Gaussian-like distribution:

$$P(x|B_{ij}^+) = \exp\left(-\sum_{k=1}^{p} w_k (B_{ijk}^+ - x_k)^2\right). \quad (3)$$

where w_k is a set of scaling factors related to the relevance of each feature k that are also learned in the process of finding the optimal point, x_{opt} (which can be seen as the prototype of the positive examples). As proposed by Maron [9] we can in fact compute K prototypes $x_{opt,k}$, where $k \in [1, K]$, using:

$$P(x_1 \vee x_2 \ldots \vee x_k \mid B_{ij}^+) = \max\left\{P(x_1 \mid B_{ij}^+), \ldots, P(x_k \mid B_{ij}^+)\right\} \quad (4)$$

We mention that the addition of each prototype increases the search space by p, therefore increasing the time and data requirements of the searching procedure.

One important observation related to our approach to detecting health events is that we have "ground truth" (nurse report or other health records) available only at a given time of the day (unknown, hence we arbitrarily chose it to be 12 pm) for which the sensor activity is not, in general, relevant. Instead, the report describes some health event that likely happened in the last 24 hours before the visit. We assume that there is at least one hour of "abnormal" sensor data during that time that reflects the reported health event. In our previous work [11, 12], we used some aggregation method (sum) over the entire 24 hour period to represent the sensor activity prior to a nursing visit. Instead, here we propose to view the previous day as a bag of 24 hours. This will make possible identification of few "bad hours" that might relate to the health event reported in the nursing visit. More importantly, it will allow for a faster detection of a possible health event, that is, at the end of each hour instead of at the end of the day.

We investigated the performance of the MIL framework using three different datasets: first obtained by simulation (see Section 3.1), second in which the ground truth was acquired by telehealth devices (Section 3.2) and the third one where the ground truth was provided by electronic medical records (Section 3.3).

2.2 Experimental Setup

Tiger Place [14, 16] is an independent living facility for seniors designed and developed as a result of collaboration between Sinclair School of Nursing, University of Missouri (MU) and Americare Systems Inc. of Sikeston, Missouri. A primary goal of Tiger Place is to help the residents not only manage their illnesses but also stay as healthy and independent as possible. Another goal of our research of equal importance was to maintain privacy of the residents. Privacy concerns were addressed using a two prong strategy. First, the MU institutional review board (IRB) performed a thorough assessment of each of our projects. Second, we pursued a collaborative design approach in our research by involving the residents in the development and testing phase of all our projects. For example, earlier focus groups with Tiger Place residents [19–21] reveal their willingness to accept non-wearable devices in their apartments while exhibiting reluctance toward wearable ones. As a consequence, and somewhat surprising for people unfamiliar with our work, we decided to deploy only non-wearable devices in Tiger Place. Another interesting outcome of our focus groups was the residents'

willingness to accept a camera sensor in their apartment as long as only the silhouette of the person is captured.

Each resident included in the study has a data logger in his or her apartment that collects data from wireless sensors (Figure 1). The data logger date-time stamps the data, and logs them into a file that is sent to a database on a secure server via a wired network connection. Forty seven networks (10 with video) have been installed in Tiger Place apartments; the video part of the network is currently under development. The sensor network consists of several types of sensors mounted in different places throughout the residents' apartments, including motion sensors, bed sensors, and stove temperature sensor. The motion sensors are placed in various places, such as bathroom, bedroom, kitchen, living room, etc.

Motion sensors are passive infrared (PIR) devices that react to the heat generated by the human body. Most PIR sensors are sensitive to hand movement up to a distance of about 10 feet, arm and upper torso movement up to 20 feet. PIR sensors send an X10 signal every 7 seconds to the data logger while human motion is present in their activity cone (typically 40° wide). We deployed a PIR sensor in each apartment room (e.g. bathroom, bedroom, kitchen, living room, etc.) of the Tiger Place residents that agree to participate in our study. The bed sensor consists of a pneumatic sensor strip mounted across the bed and a motion sensor attached to the bed headboard. The sensor strip is able to keep track of the resident's restlessness, pulse and breathing, as long as the person lies on the bed. The sensor strip and motion sensor attached to the bed are connected together and they function similarly to the motion sensors mentioned previously: they fire as long as they detect activity. The signals captured by the bed sensor (restlessness, pulse and respiration) have three or four levels of severity (low, medium, high, very high). A low pulse event is sent if the detected pulse is lower than 30 beats per minute (bpm), a normal pulse event is sent for 30–100 bpm and a high pulse event is generated at greater than 100 bpm. Similarly, a low respiration event is sent when breathing rate is lower than 6 times per minute (tmp), a normal one for 6–30 tmp and a high one for greater than 30 tmp. The four restlessness levels correspond to motion duration of 0–3 second, 3–6 seconds, 6–9 seconds, and greater than 9 seconds, respectively. We only used first restlessness level (0–3 seconds) in this paper, as the other three levels were rather infrequent in our dataset. Further technical details about the bed sensor can be found in reference [25].

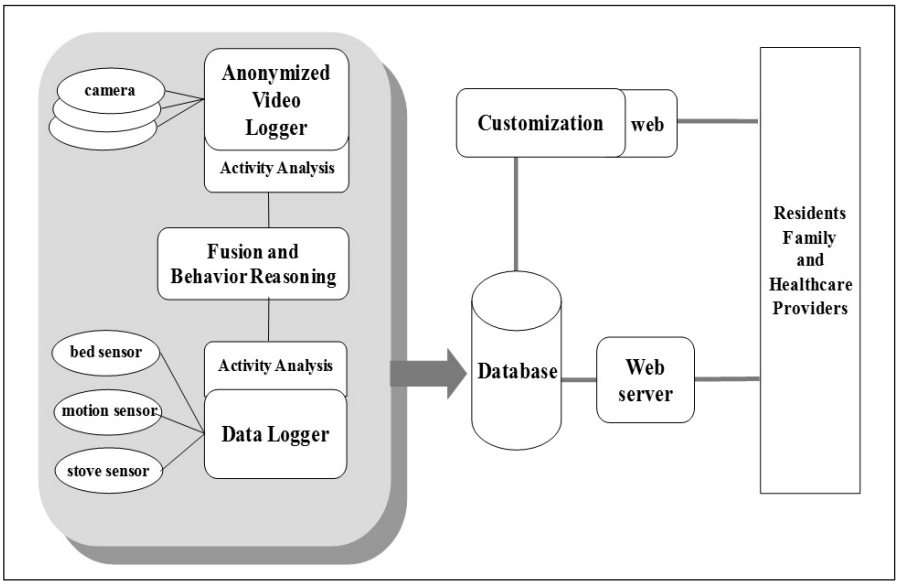


Fig. 1 The Tiger Place sensor network architecture. In each apartment, all sensor data is collected by a data logger running on a local computer. Only motion and bed sensors were used in this study. A database placed in a secure location is synchronized at midnight with all data loggers from Tiger Place. Sensor data can be viewed using a secure web interface.

As mentioned in the introduction, our early illness recognition approach is based on the intuition that if the resident does not feel well, his/her sleep and motion patterns are altered. Based on previous work [11, 12], in this study we used five features ($p = 5$) to represent the resident behavior: the total number of firings for motion, bed restlessness, low pulse and low breathing sensors, respectively, for each hour of the day before the nursing report (considered at 12 pm). The fifth feature is represented by the hour of the day when the sensor readings were made. This feature is required in order to differentiate the night time behavior from the day time one.

Although each resident lives alone in his apartment, some extra motion hits were possible due to housekeeping or occasional visits. Although our group has developed algorithms for detecting these events [16] we did not use them here since MIL should be able to automatically account for them. For example, visits are likely to occur both in negative ("feel good") and positive ("feel bad") days, hence feature vectors with an abnormally high motion values generated by a visit will be treated as "negative" instances.

The implementation of the MIL algorithm (described in Section 2) employed in this paper is summarized in ▶Figure 2.

The MIL framework recognizes when some hourly patterns appear both in the positive and negative bag and, consequently, labels them as "normal". Theoretically, only hourly patterns that appear in the positive bags should have a chance to be labeled "abnormal". In reality, we cannot guarantee that no outliers are present in the negative bags.

For example, the pattern produced by a visitor (nurse, family) while the resident is in bed during the day is likely to be considered "normal" since it might happen in both good (due to housekeeping) and bad days (due to nurse visit). In the mean time, the framework will automatically detect if there is an hour during the day when the resident is in bed only during the abnormal days."

We used the positive and negative hours from N-1 days to compute a prototype for a

Input: - ε - a threshold for deciding a positive hour
- N day labels, l_i:
$l_i = 1$ if the day was positive (feel bad)
$l_i = 0$ if the day was negative (feel OK)
- N *24 sensor data vectors, $x_k \in R^5$
For each day $i = (1, N)$
Step 1 (train).
Use data from $N-1$ days, i.e. $24*(N-1)$ data points, to compute K positive prototype, $\boldsymbol{x}_{opt,k}$, $\boldsymbol{k} = (1, K)$ and related feature weights, $\boldsymbol{w}_k \in [0,1]^P$, using equation (1);
Step 2 (test):
For each hour $j \in [0,23]$ in day i
- compute $d_j = \min_{k\in(1,K)}\{\|\boldsymbol{x}_j - \boldsymbol{x}_{opt,k}\|\}$
- if $d_j < \varepsilon$
label day i, $l_i^* = 1$ (positive);
end
End
If no positive hours found, set $l_i^* = 0$ (negative);
End
Step 3. Compare $\{l_i^*\}$ to $\{l_i\}$ for each ε to compute ROC

Fig. 2 The algorithm of the MIL implementation used in this paper

"bad" hour, x_{opt}, and the weights for each feature, $w \in [0, 1]^P$. The computing of w eliminates the need for feature normalization. However, the normalization is still performed in order to compare the MIL approach to the other classifiers. Examples of "bad" hours can be seen in ▶Figure 3. In ▶Figure 3a the resident moved a lot in his apartment around 1 am instead of being in bed. In ▶Figure 3b the resident was, uncharacteristically, in bed at 12 pm. The prototype obtained in this fashion was then used to label the hours from the N-th day. More prototypes (concepts) $x_{opt,k}$ can be obtained by adding ▶Equation 4 to the optimization procedure. We used k = 1 (denoted as MIL-1 or simply MIL) and k = 2 (denoted as MIL-2) in this paper. The optimization procedure was implemented using the fmincon function from the Matlab (http://www.mathworks.com) optimization toolbox.

An hour was labeled positive ("feel bad") using a nearest neighbor approach, that is if the distance to the positive ("bad hour") prototype, d_j, was smaller than a given threshold. When multiple prototypes were employed d_j was computed as the minimum of all distances to the individual target concepts, $||x_j - x_{opt,k}||$ where $||.||$ is the weighted Euclidean distance that uses the computed weights, w. A threshold $\varepsilon \in (0, 1)$, was used to decide if an hour was abnormal (positive) or not, i.e. if $d_j < \varepsilon$ then the hour was abnormal. Since the weights of the features were available as an output of MIL training, we used a weighted Euclidean distance. If the day had at least one positive hour it was labeled "positive", else it was labeled "negative". Multiple values of e were used for computing the ROC curves presented in Section 5. We also used area under ROC (AROC) to quantify our results. Finally, we employed a leave-one-out training-testing approach: we predicted the label for each day by training the algorithms on the rest of N – 1 labeled days.

We compared the performance of our MIL algorithm to a one class classifier (OCC) approach we used in a previous paper [12], namely an OCC-SVM, and a simple k-nearest neighbor (kNN) one. We used $k = 1$ for kNN throughout this paper.

As opposed to the regular SVM that separates the two classes in the feature space by a hyper plane, OCC- SVM [17] surrounds the target class in the feature space by a hyper-sphere. Formally, we need to minimize:

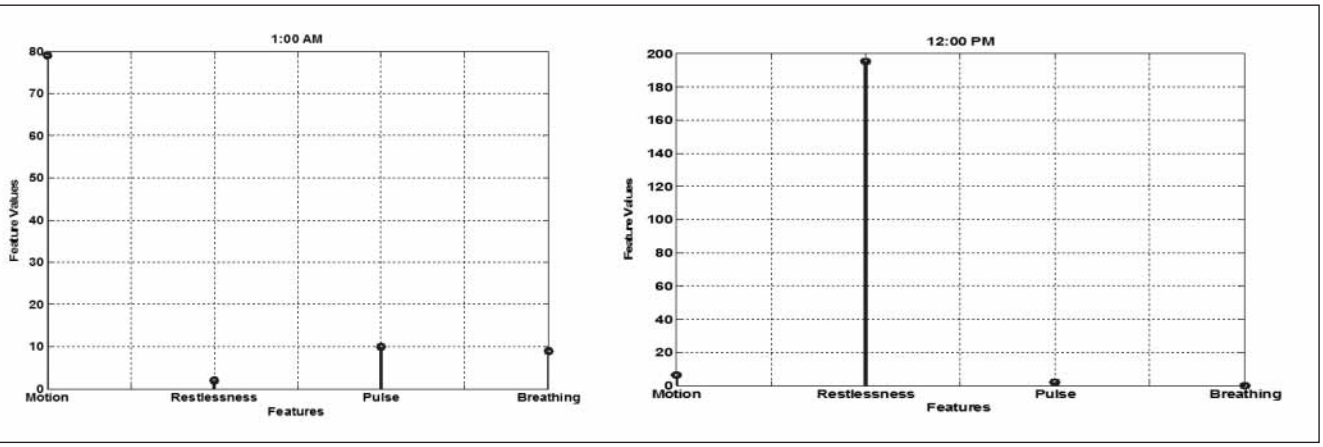


Fig. 3 Examples of "bad" (positive) hours. A "bad" night hour (a.) has high motion values (meaning that the resident away from his/her for a long time) whereas a "bad" day (b.) has high restlessness (which means that he/she is in bed during day time). Note that if the person is not in bed the entire interval of time (hour), the values for pulse restlessness and breathing are all zero.

$$R^2 + C\sum_i \xi_i , \quad (5)$$

where ξ_i are slack variables, R is the radius of the hyper-sphere and C is a constant. Also, we need a constraint that the objects be in a sphere of radius R:

$$(||x_i - a||)^2 \leq R^2 + \xi_i , \quad (6)$$

where a is the center of the sphere. In the above formulation, a and R are computed such that a percentage of the training set objects will lay inside the sphere. More details about OCCs can be found [17].

For OCCs, we used only the negative (feel good) days for training. We employed the same sliding window approach but we aggregated the sensor data in a different way: features 1–4 were the sum of sensor data for the night hours (7pm – 7am) and features 5–8 were the sum of the sensor data for the day hours (7am – 7pm). Although not necessary for the MIL approach, we normalized the features using the mean, m, and the standard deviation, s, of the $(N-1)\times 24$ feature vectors used for training, as $x_n = (x-m)/s$.

3. Datasets

3.1 Dataset 1

We generated a set of 50 days for a hypothetical Tiger Place resident using a sensor environment simulator, TigerSim [6]. TigerSim can model any Tiger Place apartment and its sensor environment. For a given resident path in the simulated apartment, TigerSim produces a list of sensor firings very similar to the one recorded by the data logger in the real environment. For dataset 1, henceforth referred to as DATA1, we simulated 50 days of "normal" and "abnormal" activity. DATA1 contained 15 "bad" and 35 "normal" days. A "normal" day consists in wake up, wash, have breakfast, read, have lunch, read, watch television and have dinner. A "bad" day would have the same activities plus time intervals with pacing behavior (walking back and forth in the living room), apartment motion during the night and lying in bed during the day. TigerSim can be downloaded from http://cirl.missouri.edu/tigersim/.

Table 1 Data for three residents from dataset DATA2. This is our telehealth ground truthed dataset, and for this reason has a low number of total records. In addition, this dataset has relative few examples of "feel bad" days.

	Total records, N	Negative (feel good) days	Positive (feel bad) days
Resident 1	66	54	12
Resident 2	69	62	7
Resident 3	27	17	10

Table 2 The data for the three residents from DATA3. As opposed to DATA2 this dataset does not represent contiguous days. In addition, two of the residents (4 and 5) and a relatively low amount of "feel bad" days.

	Total records, N	Negative (feel good) days	Positive (feel bad) days
Resident 4	441	360	81
Resident 5	744	709	35
Resident 6	499	164	335

3.2 Dataset 2

The data from dataset 2, henceforth referred to as DATA2, has been collected from three Tiger Place residents during the interval they also had a Tunstall (http://www.tunstall.co.uk) telehealth device installed in their apartment. The telehealth device recorded self administered blood pressure, pulse ox, weight measurements and answers to simple questions such as "How is your day: better than/worse than/normal?", "How is your appetite today?" and "How did you sleep last night?". The answers to those questions were used to manually label each day as "good" or "bad". The number of sensor days for each resident in DATA2 is shown in ▶Table 1. We note the time interval from DATA2 is contiguous, i.e. the data was recorded in consecutive days.

From ▶Table 1 we can observe two characteristics of this dataset. First, it has a low total number of days for each resident. This is because the telehealth equipment was only used for about four months in Tiger Place. Second, it has a small number of "feel bad" days per resident.

3.3 Dataset 3

The data from dataset 3, henceforth referred to as DATA3, consisted of sensor hits from days that span a 5-year period. The number of days for three residents considered in DATA3 is shown in ▶Table 2.

From ▶Table 2 we see that this dataset contains significantly more days than DATA2. However, this dataset is not ideal (a large amount of data and balanced number of good and bad days), since two of the residents (4 and 5) are relatively healthy and, consequently, have a relative small number of "feel bad" days.

In addition to the sensor data, we have available all the clinical records (medication, nursing visits, hospitalizations, etc) for the above three residents. The labeling of each day ("good" vs. "bad"), i.e. ground truth, was performed manually by the authors based on the nursing visit reports and other clinical records. Although we have the residents monitored for over five years now, we only included the days in which a nurse report or other clinical record was available. As a consequence, the dataset is not contiguous, that is, the sensor hits were not recorded in consecutive days.

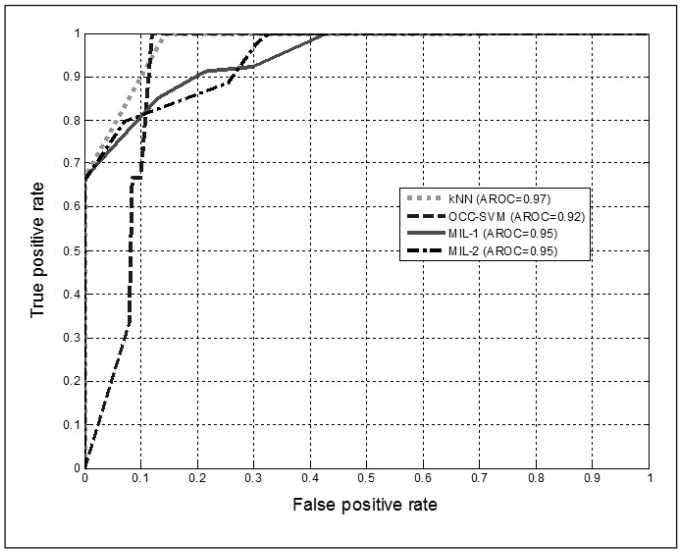


Fig. 4 Results obtained on DATA1. Both MIL variants (with one target concept, MIL-1, and with two target concepts, MIL-2) outperformed the one class SVM (OCC-SVM) classifier. The nearest neighbor (kNN) classifier was slightly better than MIL. a) Resident 1; b) Resident 2; c.) Resident 3.

The results of the illness prediction for the above datasets are presented in the next section.

4. Results

4.1 Results on DATA1

In ▶Figure 4 we show the results obtained on DATA1 dataset.

As expected, both MIL and two-class kNN outperformed the one-class SVM (OCC-SVM). kNN outperformed MIL on DATA1 due to, probably, the small size of the dataset and the low diversity of the simulated abnormal patterns. The MIL with two concepts (MIL-2) produced similar results to the MIL with one concept (MIL-1). Again, there was probably not enough diversity to learn from this dataset.

Table 3 Mean and standard deviation of the AROC results obtained on DATA2.

	Mean AROC	Std
MIL	0.727	0.035
kNN	0.653	0.023
OCC-SVM	0.590	0.040

4.2 Results on DATA2

The results on DATA2 are given in ▶Figure 5a–c.

The results are also summarized in ▶Tables 3 and 4. In ▶Table 3 we presented the mean and standard deviation of the AROC values obtained on DATA2. In ▶Table 4 we summarized the true positive values obtained at a false positive rate of 0.3 for the same dataset.

From ▶Tables 3 and 4 we see that, for DATA2, MIL outperformed both kNN and OCC-SVM (difference significant at $p = 0.05$). Since DATA2 represents a contiguous period of time of about one to two months, the abnormal behavioral patterns are probably not very diverse. However, this dataset is representative for the deployment scenario of our early illness detection method: each day, the classfier will be trained employing a MIL approach on sensor data and health events from the last one or two months. We note that much longer training intervals, while beneficial from the classifier point of view, might not be appropriate due to a gradual change in behavior caused by aging.

	MIL	OCC-SVM	kNN
Resident 1	0.72	0.58	0.40
Resident 2	0.62	0.53	0.34
Resident 3	0.56	0.54	0.50
Mean ± std	0.63 ± 0.08	0.55 ± 0.03	0.41 ± 0.08

Table 4 True positive (detection) rate at 0.3 false positive rate obtained on DATA2

4.3 Results on DATA3

The results on DATA3 are presented in Figure 5a–c.

The results are also summarized in ▶Tables 5 and 6. In ▶Table 5 we presented the mean and standard deviation of the AROC values, while in ▶Table 6 we summarized the true positive values obtained at a false positive rate of 0.3 obtained on DATA3.

From ▶Tables 5 and 6 we see that both two-class classifiers, MIL and kNN, outscored the OCC classifier (OCC-SVM). However, for this dataset there is no difference in performance between MIL with one prototype, MIL-1, and kNN. The reason might be the more diverse behavior patterns present in this dataset. This has been confirmed by the fact that when two prototypes were used (see MIL-2 results) MIL results drastically improved and outperformed kNN. We note that MIL performed consistently in both real datasets, with an average AROC of about 0.7 when a prototype was used and about 0.76 when two concepts were employed. While the values themselves are modest, we believe that it proves the utility of the MIL approach.

5. Discussion

In this paper, to test out multiple learning (MIL) approach, we employed three datasets that are typical for most eledercare applications. Each dataset type has its strength and weaknesses. The simulator dataset, DATA1, has the advantage that can produce ground-truthed balanced data of any size. However, it is relatively hard to simulate realistic behavior patterns. Using telehealth devices (as in DATA2) would cre-

ate a great opportunity for attaching ground truth to sensor data. However, the use of telehealth devices is challanging in elderly. For example, some residents positioned the blood pressure cuff incorrectly, resulting in erroneous readings that falsely concerned the nursing personnel. Others residents had difficulty navigating the user interface of the telehealth hub/modem (that was lacking both a touch screen and a voice interface) resulting in many missing values and early termination. We plan to report a complete account of our findings related to the use of telehealth devices by elderly in a future paper. Using medical records to ground-truth sensor data (as in DATA3) is difficult unless the electronic health record (EHR) is integrated with the sensor network. In Tiger Place we are currently developing an EHR [23] that integrates medical records with sensor data. This new EHR will create the possibility of automatic extraction (as opposed to our manual approach) and association of medical and sensor data.

In general, two class classifiers outperform their one-class counterparts and they should be used whenever possible. As we see in ▶Figure 4, when plenty of data is available to cover all possible patterns (here due to the limitation of our simulator script) kNN type classifiers are very powerful. However, in our case, three factors make using a two class classifier difficult: health events are rare, the ground truth is uncertain due to the delay between the event and its recording in the medical record and data cannot be used across patients. MIL addresses only the second problem. MIL addresses ground truth temporal uncertainty since we cannot tell exactly when the abnormal behavior happened. Another advantage of MIL is that it offers the possibility to detect the abnormal patterns on a hourly basis while the other approaches operate at day level. This is possible because our target concept is an abnormal hour, hence we can verify if each hour is abnormal or not. Essentially the original MIL variant [9] used in this paper is a one class classifier. We belive that a stronger classfier such as a SVM trained with a MIL framework can produce better results. Moreover, the current MIL framework does not account for the fact that a day or an instance are not exclusively "good" or "bad" but they have certain memberhip degree in both categories. In [22] we proposed a fuzzy MIL framework, FUMIL, that intends to address the membership degree problem.

Table 5 Mean and standard deviation of the AROC results obtained on DATA3

	Mean AROC	Std
MIL-1	0.700	0.050
MIL-2	0.760	0.090
kNN	0.703	0.045
OCC-SVM	0.590	0.041

As observed from ▶Figure 6 and ▶Tables 5 and 6, using multiple prototypes (2 in this case) can greatly improve the performance of the MIL framework. However, an open problem, similar to the cluster validity issue from clustering, is how to choose the number of prototypes.

Our data can be seen as a time series that, consequently, could be modeled using hidden Markov Models (HMMs). There are two potential issues with using HMMs for our problem. First, the sequence length is variable and unknown. Unless the condition is specified, one can not choose the onset length of the medical condition. The onset time interval can vary from several hours (a case of indigestion) to several weeks or months. The second HMM problem is that in some cases we have only 10 examples of "bad days" which are not enough for training an HMM, since we can not assume a left-to-right model. We believe that in order to use an HMM approach one has to target just one condition where much is known about onset time in order to have an idea of the HMM sequence length and have plenty of training data. In our case, the "don't feel well" assessment is a judgment call based on nursing visit notes, hence non specific. Our approach is targeting acute conditions with about 1 day onset time. While it is technically possible to design an HMM and train it with sequences of length 24 (number of hours in a day) our exploratory results were about 20% lower than the MIL and kNN approach.

The results obtained using MIL for six Tiger Place residents, an average AROC of 0.7, are promising. However, our study has several limitations. First, the sample size and data sets were small. Second, the labeling of the data (ground truth) was not very reliable when medical records were used (DATA3). We hoped to solve this problem by employing telehealth devices (as we did in DATA2). However, we were forced to discontinue the use of our telehealth devices due to problems linked to incorrect utilization by some residents, as previously mentioned. Third, other factors aside of health conditions (such as exciting news) may influence the sensor readings. This might lead to unwanted false alarms that might annoy the nursing personnel. To reduce the number of false alarms, we plan to combine the proposed algorithm (i.e. decision level fusion) with other approaches and sensor systems, such as video and radar, which we have recently deployed in Tiger Place.

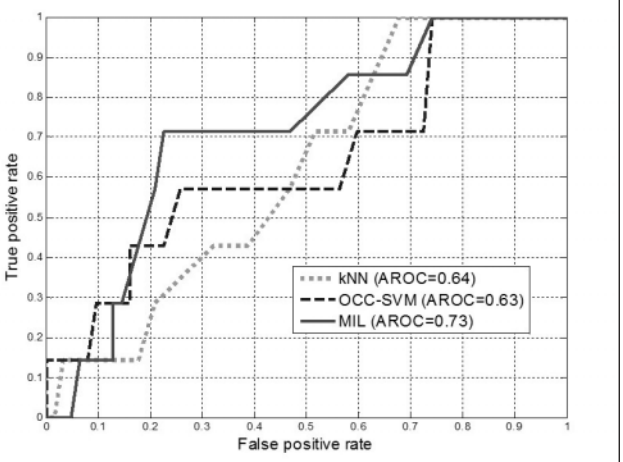
6. Conclusions

We described a method for detecting early signs of illness in elderly residents of Tiger Place based on unobtrusive monitoring sensors. The detection of early signs of illness may help nursing staff provide interventions that might prevent grave clinical events such as heart attacks or strokes. We have also shown how simulated sensor data can be used in algorithm development and testing.

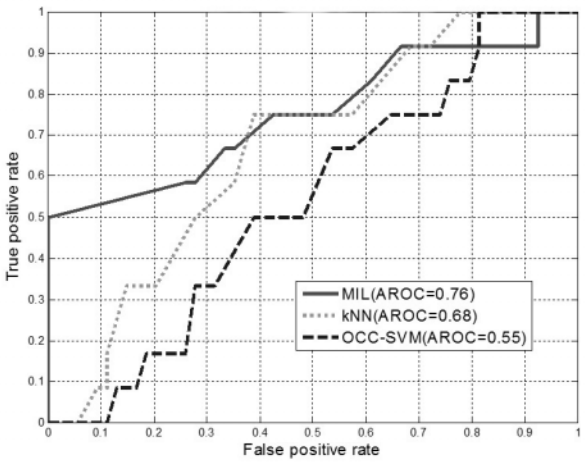
Finally, we acknowledge that our MIL testing in this paper is limited. However, in

Table 6 True positive (detection) rate at 0.3 false positive rate obtained on DATA3

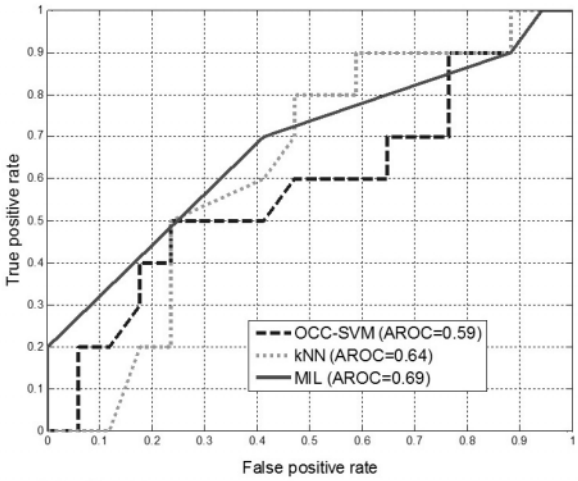
	MIL-1	MIL-2	OCC-SVM	kNN
Resident 1	0.54	0.62	0.4	0.69
Resident 2	0.63	0.80	0.35	0.72
Resident 3	0.60	0.84	0.4	0.50
Mean ± std	0.59 ± 0.05	0.75 ± 0.11	0.38 ± 0.03	0.64 ± 0.08



a) Resident 1

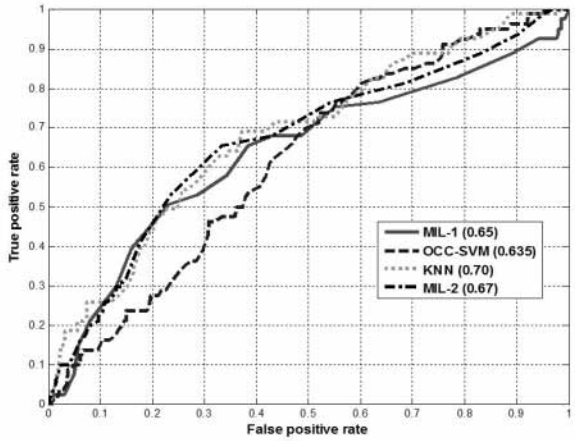


b) Resident 2

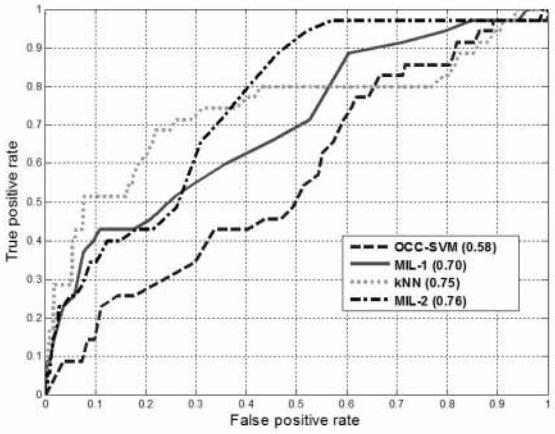


c) Resident 3

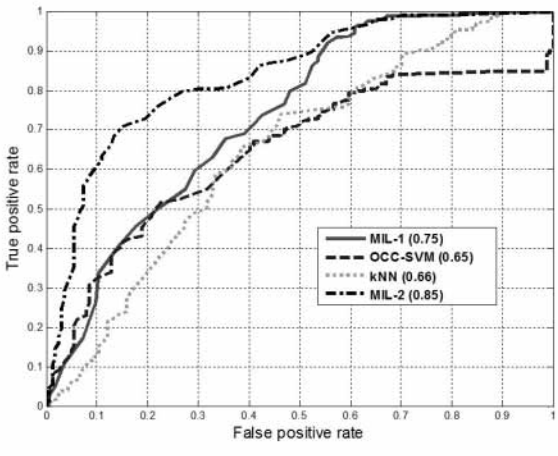
Fig. 5 Results on DATA2. In this dataset, MIL with one target concept (MIL) outperformed both nearest neghbor (kNN) and one class SVM (OCC-SVM) classifiers. a) Resident 4; b) Resident 5; c.) Resident 6.



a) Resident 4



b) Resident 5



c) Resident 6

Fig. 6 Results obtained on DATA3. kNN was the best classifier for Resident 5 while MIL with 2 target concepts (MIL-2) was the best classifier for Resident 6. Clearly, the effectiveness of a classifier depends on the type of the dataset.

Tiger Place we have developed a real time framework for testing early illness classifiers based on nursing feedback [24, 25]. A proper testing of the proposed MIL based classifiers will be possible only by their deployment in our real time validation framework. Moreover, the deployment of the Tiger Place EHR [22] will make it possible for further integration of the medical and sensor data.

References

1. Andrews S, Tsochantaridis I, Hofmann T. Support vector machines for multiple-instance learning. Advances NIPS; 2003. pp561–568.
2. Chen JP, Reich L, Chung H. Anxiety disorders. West J Med 2002; 176: 249–253.
3. Cuddihy P, Weisenberg J, Graichen C, Ganesh M. 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. Dietterich TG, Lathrop RH, Lozano-Perez T. Solving the Multiple-Instance Problem with Axis-Parallel Rectangles. Artificial Intelligence Journal 1997; 89: 31–71.
5. Gale CK, Millichamp J. Generalised anxiety disorder. Clinical Evidence 2011; 10: 1002.
6. Godsey C, Skubic M. Using Elements of Game Engine Architecture to Simulate Sensor Networks for ElderCare. Proc 31st Ann Int Conf of the IEEE EMBS; Minneapolis, MN; Sept. 2–6, 2009. pp 6143–6146.
7. Hayes TL, Pavel M, Kaye JA. 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.
8. Maron O, Lozano-Pérez T. A framework for multiple-instance learning. Proc of the 1997 Conf on ANIPS 1998; (10): 570–576.
9. Maron O, Ratan AL. Multiple-Instance Learning for Natural Scene Classification. Proc 15th ICML, 1998. pp 341–349.
10. Morris M, Intille S, Beaudin JS. Embedded assessment: overcoming barriers to early detection with pervasive computing. In: Proc of PERVASIVE 2005. Gellersen GW, Want R, Schmidt A, eds. Springer-Verlag; 2005. pp 333–346.
11. Popescu M, Florea E, Skubic M, and Rantz M. Prediction of Elevated Pulse Pressure in Elderly Using In-Home Monitoring Sensors: A Pilot Study. 4th IET Int Conf on Int Env; Seattle; July 21–22, 2008.
12. Popescu M, Skubic M, Rantz M. Predicting abnormal clinical events using non-wearable sensors in elderly. AMIA Fall Symp; San Francisco, CA; 2009 Nov 14–18. p 1010.
13. Popescu M, Mahnot A. Early illness recognition in older adults using in-home monitoring sensors and multiple instance learning. IDAMAP-10; Washington DC; Nov 12, 2010. pp 21–25.
14. Rantz MJ, Marek KD, Aud MA, Johnson RA, Otto D, Porter R. Tiger Place: A New Future for Older Adults. J of Nursing Care Quality. 2005; 20(1):1–4.
15. Rowan J, Mynatt ED. Digital Family Portrait field trial: support for Aging in Place". Proc SIGCHI Conf on Human Factors in Comp. Sys. New York: ACM Press, 2005. pp 521–530,
16. Skubic M, Alexander G, Popescu M, Rantz M, Keller J. A Smart Home Application to Eldercare: Current Status and Lessons Learned. Tech and Health Care 2009; 17 (3): 183–201.
17. Tax DMJ. One-class classification. PhD Thesis; TU Delft, NL; 2001.
18. Zhang Q, Goldman SA, Yu W, Fritts JE. Content-Based Image Retrieval Using Multiple-Instance Learning. Proc of the 19th ICML; July 2002. pp 682–689.
19. Demiris G, Rantz MJ, Aud MA, Marek KD, Tyrer HW, Skubic M, and Hussam AA. Seniors' Attitudes Towards Home-based Assistive Technologies. 29th Annual MNRS Research Conference; Cincinnati, Ohio, April 1–4, 2005.
20. Demiris G, Skubic M, Rantz MJ, Harris K, Hensel B, Aud MA, Lee J, Burks K, Oliver DR, He Z, Tyrer HW & Keller J. Older Adults' Attitudes Towards Smart Home Features. International Conference on Aging, Disability and Independence (ICADI); St Petersburg, Florida; February, 2006.
21. Demiris G, Skubic M, Rantz M, Hensel B. Smart Home Sensors for Aging in Place: Older Adults' Attitudes and Willingness to Adopt. The Gerontologist 2006; 46 (Special Issue 1): 430.
22. Mahnot A, Popescu M, "FUMIL-Fuzzy Multiple Instance Learning for Early Illness Recognition in Older Adults. To appear in Proc IEEE WCCI; Brisbane, Australia; 2012.
23. Rantz MJ, Skubic M, Koopman RJ, Alexander G, Phillips L, Musterman KI, Back JR, Aud MA, Galambos C, Guevara RD, Miller SJ. Automated technology to speed recognition of signs of illness in older adults. In press – 2012, Journal Gerontological Nursing. 2012; 38 (4): 18–23.
24. Popescu M, Chronis G, Ohol R, Skubic M, Rantz M. An Eldercare Electronic Health Record System for Predictive Health Assessment. IEEE 13th International Conference on e-Health Networking, Applications and Services; Columbia, MO; June 13–15, 2011. pp 194–196.
25. Skubic M, Guevara RD, Rantz M. Testing classifiers for embedded health assessment. To appear in Proc of Intl Conf on Smart Homes and Health Telematics; June 12–15, 2012. Italy.
26. Mack, DC, M. Alwan, and B. Turner. A Passive and Portable System for Monitoring Heart Rate and Detecting Sleep Apnea and Arousals: Preliminary validation. Proc of the Transdisciplinary Conference on Distributed Diagnosis and Home Healthcare 2006 D2H2; Arlington, VA; 2006. pp 51–54.