Multilabel Classification of Nursing Activities in a Realistic Scenario

Farina Faiz, Yoshinori Ideno, Hiromichi Iwasaki, Yoko Muroi, Sozo Inoue

Abstract Nursing activities are complex in nature: in real scenario, when nurses perform care-giving activities, they can perform concurrent activities or multiple activities at a time. With traditional supervised learning it is very difficult to classify them especially when there is no proper information of sequence and duration of activities performed. In this paper, we used a new nursing dataset collected in complete natural setting. As our observer could not enter the room of patient to record the duration of each activity, we utilized multilabel classification in this scenario. We considered our work both as multiclass classification and multilabel classification. We applied 1) Random Forest for multiclass classification, 2) Binary Relevance and 3) Random Forest SRC (Survival, Regression and Classification) for multilabel classification. By evaluating the result, we found out that Random Forest SRC achieved highest performance that is around 86% precision, 62% recall and 63% F1 score whereas Random Forest achieved around 10% precision, 11% recall and 10% F1 score and Binary Relevance achieved around 28% precision, 13% recall and 17% F1 score. We also explained why multilabel classification is necessary to recognize activities in real scenario problems.

Farina Faiz
Kyushu Institute of Technology, e-mail: farinashama@gmail.com

Yoshinori Ideno
Global business division, CARECOM CO.,LTD, e-mail: y_ideno@carecom.co.jp

Hiromichi Iwasaki
Dept of Infection Control and Prevention, University of Fukui Hospital, e-mail: hiwasaki@u-fukui.ac.jp

Yoko Muroi
Dept of Infection Control and Prevention, University of Fukui Hospital, e-mail: ymuroi@u-Fukui.ac.jp

Sozo Inoue
Kyushu Institute of Technology, e-mail: sozo@brain.kyutech.ac.jp

1 Introduction

HAR (Human Activity Recognition) has always been a challenging and crucial problem especially in realistic scenario. A huge amount of researches have been conducted due to the exceptional development of microelectronics and computer systems and availability of wearable sensors and mobile devices that have small size, high computational power and low cost [1]. Together with human activity recognition, the number of researches has grown on nursing activity recognition as well. In order to provide standardized and efficient health care to elderly people, nursing activity recognition is necessary. It can lead to better understanding and improvements in medical care [2]. The application of efficient nursing activities leads to two important research questions. 1) Can a system recognize activities in realistic scenario? 2) What if the nurses performing multiple activities at a time?

Many researches were being conducted to recognize the activities of nurses [2-5]. Most of them were performed in a controlled environment except [2, 3]. The problem with the controlled environment is that data size is smaller, subjects are bound to perform limited activities whereas in real life these activities are performed in a quite different manner. We can consider the real scenario of a nursing care center where nurses can perform multiple activities at a time. Moreover, in real life there is no restriction or proper sequence of performing those activities: nurses are free here to work in their own way. There are some activities in nursing that sometimes take long time, sometimes take short time to finish. Along with those activities, nurses can do some other tasks that take short time but are very crucial ones. Most of the time these activities could not be predicted with traditional machine learning algorithms even if we know that those activities still exist. I order to solve this issue, multilabel classification can be applied. Although multilabel classification is originated from text categorization problem, it is increasingly required by modern applications such as protein function classification, music categorization and semantic scene classification [6]. In text classification problem each document can belong to multiple topics simultaneously. Similarly, in activity recognition we can consider that within specific time window, an instance might have multiple labels. If we want to solve such problem, we need a suitable dataset to conduct such research. Even if the goal here is very clear that we want to recognize those concurrent or multiple activities of nurses, there is no such dataset that goes with our requirement. That is why the number of works regarding this field is rarely considered. However, it is true that with the advancement of assistive healthcare technologies the demand of recognizing nursing activities is increasing. Nursing activities are complex in nature and in real it can be concurrent as well. Taking these issues into account multilabel classification should be considered in this aspect.

Based on our study, currently there are no such works that utilize multilabel classification on realistic nursing dataset. In this paper we introduced a new realistic scenario dataset that was collected from a University Hospital. It contains nursing activities performed by different nurses on separate days. We utilized both multiclass and multilabel classification methods in order to recognize the activities of nurses. The reason to consider this scenario as a multilabel problem is while recording and

annotating the label of nurses our observer could not enter the patient's room. So they were recording activities after certain time. The purpose is to utilize multilabel classification methods in a smarter way and analyze the results by which we can handle such challenging dataset.

In this paper, we conducted some initial preprocessing on raw sensor data from which we extracted some statistical and motion features. Then we adapted our dataset to fit into multiclass and multilabel classification. This process was very important part as we were annotating labels as per the structure required by each method. After that, we applied Random Forest for multiclass classification and for multilabel classification we used Binary Relevance and Random Forest SRC [7]. We compared the results achieved by each method. Random Forest SRC achieved around 86% precision, 62% recall and 63% F1 score in our dataset whereas Random Forest achieved around 10% precision, 11% recall and 10% F1 score and Binary Relevance achieved around 28% precision, 13% recall and 17% F1 score. Finally, we discussed some details regarding those findings and provide future directions related to our work.

The paper is organized as follows: Section 1 covers the introduction of the paper. In Section 2, we present some related works on nursing activity recognition and multilabel classification. Section 3 presents dataset overview and preprocessing. In Section 4, we present the methodology of our entire work. In section 5 we present the results of our experiments and evaluation we conducted for the problem setting. In Section 6, we present the discussion and motivation behind this work. Finally, we conclude the paper with some future work points in Section 7.

2 Related Work

Human activity recognition is one of the most promising research areas these days. With increasing and availability of body-worn, ambient and object sensors, the sensor-based human activity recognition is growing and attracting attention in a number of disciplines and application domain [8]. Over the past decade sensor technologies, especially low-power, low-cost, high-capacity and miniaturized sensors, wired and wireless communication networks made substantial progress [9][10]. Researchers surveyed that in healthcare monitoring wearable sensors are widely used and reviewed the state-of-the-art in this field[11]. In recent years a number of researches have been conducted to recognize daily human activities using body-worn sensors or accelerometers[12-14]. In these works, they used supervised machine learning and deep learning methods to learn simple activities like standing, sitting, walking, jogging etc. Not only daily living activities but also researches have been done on nursing activities in health care centers[4]. The aim of the work was to build a recognition system for nursing activities. However, the limitation of the study was the activities were performed for a short time. In order to control infection, work has been also done to recognize nursing activity. They used five tri-axial accelerometers attached to nurse's body and recognized six activities [5]. The common thing

among these researches is that they used somehow well-structured dataset which were collected in a controlled environment instead of natural settings. Even though they considered those settings as practical nursing environment, the activities were either pre-defined or performed for a very short period. Another research has been conducted considering big nursing data in realistic scenario. They collected a real nursing dataset for mobile activity recognition that can be used for supervised machine learning and proposed a method for recognizing activities for an entire day utilizing prior knowledge about the activity segments in a day [2]. In this case they were known about the information regarding duration of each performed activity.

It is true that a number of low power consumption and cheaper sensors are available these days. Advancement in ubiquitous and pervasive computing resulted in those sensors. It is really very difficult to infer meaningful information form sensor data. In realistic scenario a human can perform activities concurrently or complex activity simultaneously in different combination [10]. The fact is inferring such activities from sensor readings are challenging. Based on this fact, researchers investigated the suitability of multilabel classification inspired by decision trees as a proposed solution [15]. As multilabel classification is very common in text and image classification; research work has also been conducted to localize human activity from video sequence. Their approach proposes a matrix completion approach to the problem of WSL for multi-label learning for video [16]. There is another work where authors proposed a Bayesian framework for multilabel classification in completely realistic setting. They collected the labels from the participants who were using sampling application running on mobile phone [17]. The problem of the study was most of the users labelled their activities when they were free. So, the dataset was unbalanced and had no varieties among labels.

The related studies mentioned above are considering simple repetitive activities for classification. Despite being complex activities in nature these problems are being solved in traditional multiclass classification manner which might fail to infer simultaneously performed activities. Some of the research work attempted to address the multilabel issue in different sensor positions [18]. Multilabel classification techniques were also used to solve simultaneously occurred activity recognition problem. In this work they exploited the temporal relations between the subsequent activities as they had complete information of activity duration [19]. For the problem setting where temporal relation is unknown, their technique might not bring promising result.

With a complete realistic scenario where there is no restriction of performing any activity, no prior knowledge of activity duration or sequence, it becomes a great challenge to handle them in traditional way. To our best knowledge, there is no prior work which has explicitly addressed the multilabel classification specially to recognize nursing activities in a complete natural setting based on only two body-worn sensors. There is an existing study which is considering real nursing data for their activity recognition but not considering the problem as a multilabel[2]. Because in that case the duration of each activity was known. In our case, it is a multilabel problem in consideration to the natural way of nursing where nurses keep doing multiple works at a time and also observers could not enter the room of patients.

Moreover, we do not have any particular information of the sequence and duration of activities. So, it is very much possible in a certain time that the nurses performed activities in concurrent or sequential manner.

3 Dataset Overview and Preprocessing

The section provides an overview of dataset we used. The data was collected in a University Hospital. In this section, we will describe the sensors, nursing activities selected for this dataset. For a clearer understanding we also provide the properties of the dataset in Table 3.

Nursing Activities: A number of activities performed by nurses was recorded in hospital. The speciality of this dataset is that it was collected in a real-life setting. There was no control of performing activities for nurses. They were performing each activity according to the conditions of patient. That is why the number and nature of activities are very different from each other. The activity classes are not like regular activities in a sense that we collected data focusing on activities related to hand hygiene. We are considering these activities as multilabel because we could not record the exact timing in patient rooms. For this experiment we used 30 different activities. Table 1 shows the list of those activities. Figure 1 represents the frequency activities by all nurses.

Sensors: Position sensors, absolute pressure sensors were used to collect data. Each nurse was carrying a container of sanitizer where a locating sensor was attached and it was capturing two-dimensional position (x, y). Each of the nurse was carrying an absolute pressure sensor in their chest pocket that gives highly accurate barometric pressure measurement and temperature. A reference machine was installed at 1.3 meter from floor in each nursing ward. It was used to calculate the pressure as the sea level pressure or in Pa. Nurses walked for five minutes carrying two sensors. The position and pressure data were collected with a sampling rate of 2 Hz.

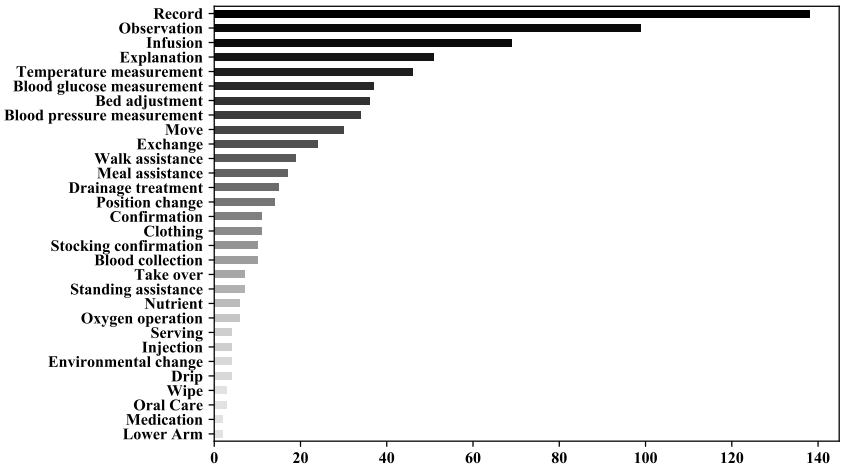

Figure 1: Activity records count in nursing dataset

Data Selection: During the data collection period, data from 12 nurses were collected for 12 days. Each nurse was performing activities for one day from around 9 am to 4 pm. The duration and number of activities completed by them were fully different from each other. Some activities were common among them but many of the activities were different that makes this dataset more diverse and challenging to be handled. Due to the convenience and maintenance of similar activities, we selected data of 4 nurses for our experiments. The reason was most of the nurse's activities were not matching with each other and many of the them were performed only once or twice that can make the dataset highly imbalanced. Another issue was the absence of particular location and pressure data. Therefore, we selected those nurses who have appropriate data and performed similar activities.

Data Preprocessing: Before fitting the data into machine learning methods we took the raw signals and applied filtering such as particular nurse's location and pressure value by cross checking the duration of action history provided. The raw signals contain hex value of pressure and temperature that we needed to convert. There were some missing values; for the convenience of our work those were removed. After getting x, y, pressure and temperature value, we split the whole into 60 second time windows. The reason is nursing activities were complex in nature and a larger window is good to consider for recognizing such activities. Additionally, we checked the action history of each nurses and observed that most of the activities including single or multiple actions took one minute on average. We repeated this same procedure to preprocess individual nurse's data.

Label Annotation: The label annotation or action sequence of this dataset is as same as the traditional HAR dataset. In traditional HAR data we have one target class for each instance. In this dataset, nurses performed multiple activities from one or five minutes. For example, a nurse at first recorded patient's condition, then observed him/her, measures temperature or blood pressure etc. One or two observer were always there to observe and write down the action sequences done by nurses.

Figure 2 shows that the dataset contains both single and multiple labels. The plot represents the total instances used with label distribution for our experiment.

Table 1 The list of activity classes

No.	Activity class	No.	Activity class
1	Record	16	Confirmation
2	Observation	17	Stocking confirmation
3	Infusion	18	Blood collection
4	Explanation	19	Take over
5	Temperature measurement	20	Standing assistance
6	Blood glucose measurement	21	Oxygen operation
7	Bed adjustment	22	Nutrient
8	Blood pressure measurement	23	Environmental change
9	Move	24	Drip
10	Exchange	25	Injection
11	Walk assistance	26	Serving
12	Meal assistance	27	Oral care
13	Drainage treatment	28	Wipe
14	Position change	29	Medication
15	Clothing	30	Lower arm

Feature Extraction: We extracted 12 features from the timeseries data. In order to extract the features, we first defined the window size of timeseries data. We set 60 second window size to extract those 12 features. The window size is comparatively larger because the activities we are handling in this dataset are complex. Activities such as walking, jogging or running that involve the complete body or several parts are more easily recognized and also permit one to optimize the window duration. However, larger window size is good for recognizing complex activities[20]. Another reason is the duration of each performed activities was minimum 60 seconds. Therefore, a smaller window size has the possibility of information loss. The scope of this paper is to recognize such complex activities with larger window size rather than evaluating optimized window. The features we extracted for our experiment are maximum, minimum, mean and standard deviation of x-axis, y-axis position, pressure and temperature values. Besides, we calculated the velocity and acceleration from the location sensors x-axis and y-axis. These two features are playing a vital role in such case because we did not have accelerometer data. Only with location it is helpful to find out what the velocity or acceleration of nurses when they were working in different locations.In Table 2 the list of features is shown. Minimum and maximum values are computed for location, pressure and temperature. Min and Max provide what were the minimum and maximum values of those variables from sensor signals.

The mean or average summarizes the values present in a particular timeseries data. Standard deviation is the summary measure of the differences of each datapoint from the mean. If $X = x_1, x_2, x_3 ..... x_n$are the timeseries data, so the mean is

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

Table 2 Features extracted from the raw data

No.	Features	No.	Features
1	X_{Min}	7	Y_{Mean}
2	X_{Max}	8	Y_{Std}
3	X_{Mean}	9	P_{Mean}
4	X_{Std}	10	T_{Mean}
5	Y_{Min}	11	Velocity
6	Y_{Max}	12	Acceleration

Table 3 Properties of Dataset

Properties	Description
Total Subject	12
Subject for experiment	4
Sensors	Pressure and position
Duration	1 day per subject
Sampling rate	2Hz
Activities	30
Features	12

and the standard deviation will be

$$s = \sqrt{\frac{\sum (x_i - \mu)^2}{n}} \tag{2}$$

The velocity and acceleration were calculated based on x and y values. If we consider at time t_i the location of nurse is x_i and y_i and at time t_n the location of nurse is x_n and y_n so we find the velocity

$$v = \frac{\sqrt{(x_i - x_n)^2 + (y_i - y_n)^2}}{(t_n - t_i)} \tag{3}$$

We find acceleration as

$$a = \frac{v}{(t_n - t_i)} \tag{4}$$

.

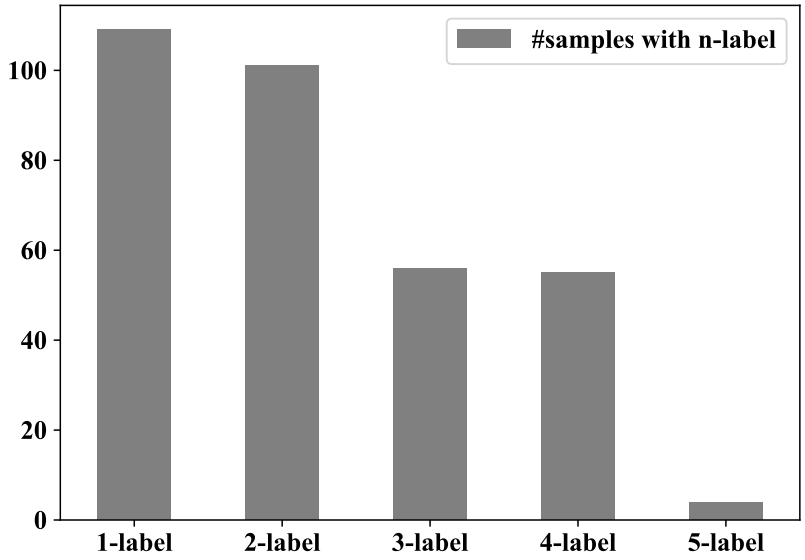

Figure 2: Label distribution

4 Methodology

Our methodology considers this problem both as a multiclass and multilabel classification problem. As the dataset contains both single and multiple label we need to first make the dataset adaptive for both multiclass and multilabel algorithm. In Figure 2 we showed the flow diagram we consider for nursing activity recognition.

Multiclass Classification: A multiclass classification task assumes that one instance has one and only one label. Most of the human activity recognition task is considered as a multiclass classification. However, in our dataset a large number of samples has multiple labels. For a proper evaluation we converted this multilabel dataset into multiclass. One of the most challenging part of this dataset is that we do not know the exact duration of activities performed. Even though the dataset contains one activity with one-minute duration, it also includes samples with multiple labels with one-minute duration. Therefore, we cannot assume an average duration of one activity. We could not even discard all the other labels keeping only the first one because it will no longer have a perfect ground truth. In order to overcome this issue, we considered a sample having multiple labels as an individual by labelling them with each of their labels. Figure 4 shows how labels are assigned for multilabel and multiclass problem. We adapted a multilabel dataset into a multiclass dataset by assigning each label to the same instance if it has multiple labels.For example X is the set of instances where $X = x_i, x_2, x_3 ..... x_n$ and Y is the set of labels where $Y = y_i, y_2, y_3 ..... y_n$. For some instance $x_i \rightarrow y_i$ where $y_i \in y_1, y_2, y_3 ..... y_n$,we as-

signed each of the labels in the set one by one for the instance x_i. Thus, this process making it as a multiclass dataset without losing the information of ground truth.

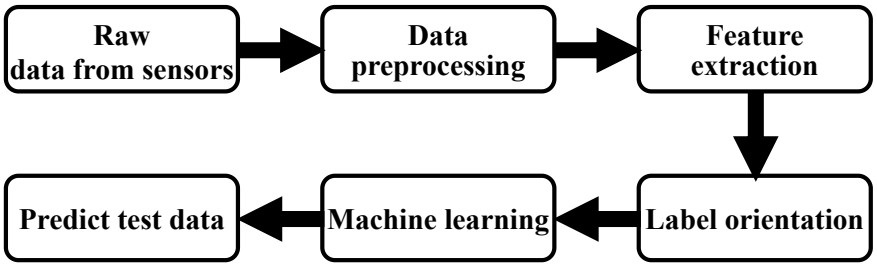

Figure 3: Flow chart for multiclass and multilabel classification

Multilabel classification: Unlike multiclass classification in multilabel classification each sample is assigned to a set of labels. In our dataset, the nurses have performed a series of activities for a certain duration. It is not clear that which activity was performed for how long. Because the observers could not enter the room to record each activity. Therefore, we considered it as a multilabel problem. Besides, in real scenario while providing nursing care to patients it is highly probable that they are doing multiple activities at a time such as recording a patient's data and observing, providing meal assistance, medication etc. These activities look similar and are performed interchangeably even though they are different. With traditional machine learning methods, they are less probably being recognized. Multilabel classification algorithms can perform better in such cases.

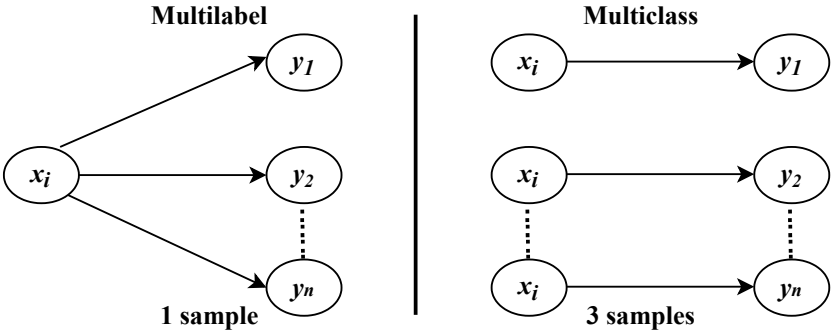

Figure 4: Label assignment in multiclass and multilabel problem

Classification: We used three types of algorithms for the classification tasks. For traditional multiclass classification we used Random Forest classifier. Random forest is one of the finest methods that combines a number of decision trees trained over

an equally distributed subsets of data. We set 100 estimators for our data. For multilabel classification we used binary relevance; a well-known problem transformation method to solve multilabel problem. The task of binary relevance is to transform the multilabel problem into a single-label classification. It is a very computationally efficient method in multilabel prediction. Another efficient algorithm is Multilabel Random Forest for solving multilabel problems. This one comes in mlr package of R. The core difference from standard random forest is that it uses normalized Gini Index splitting rule [21]. Labels were set as binary values (0 or 1) and logical factors (True, False) respectively for Binary relevance and RandomForestSRC.

Cross Validation: We used 5, 10, 15 and 20-fold cross-validation for each algorithm. As we considered four nurses data, we conducted our experiments in two ways. Firstly, we trained the models with individual nurse data; in this case we used the above mentioned cross-validation just to get a better use of all data and the to make to better evaluate the performance of model. Secondly, we trained our data on nurse 1, 2 and 3 and tested it on nurse 4. The reason is that the nurse's activities are very different from each other. Moreover, the way and duration of performing same activity by all nurses are also different. We trained our model in different cross validation settings. Then we compared the results of each methods and summarized them.

Evaluation Metrics: We chose four evaluation metrics in order to analyze the performance of applied method in our dataset. For classification method we calculated the accuracy and weighted precision, recall and f1 score. For multilabel classification we used subset accuracy, weighted precision, recall and f1 score. As for weighted average precision, recall and f1 score we considered the weighted one among instead of micro and macro. The reason is weighted calculate metrics for each label, and find their average weighted by support or the number of true instances for each label. Our focus is to evaluate the performance of both multiclass and multilabel with equivalent metrics. Let T be the total instances, N be the number of total labels. Let w_i be the weight for i^{th} label, then we find the precision p_i. Similarly, we find the recall and F1 score r_i, f_i accordingly. As we know that precision, recall and F1 score can be calculated as *TP/(TP+FP)*, *TP/(TP+FN)*, *2*Precision*Recall/(Precision+Recall)*, we can calculate

$$Precision_{weighted} = \frac{\sum_{i=1}^{N} w_i p_i}{T} \tag{5}$$

$$Recall_{weighted} = \frac{\sum_{i=1}^{N} w_i r_i}{T} \tag{6}$$

$$F1score_{weighted} = \frac{\sum_{i=1}^{N} w_i f_i}{T} \tag{7}$$

For multiclass accuracy we calculated the average number of correct predictions. Subset accuracy is the strictest metric indicating the percentage of samples those have all their labels classified correctly. If Y and Z are the sets of true and predicted labels, then subset accuracy is

$$Subset_{accuracy} = \frac{\sum_{i=1}^{T} [Y_i = Z_i]}{T} \quad (8)$$

5 Experimental Results and Evaluation

In this section we going to show the results achieved from our experiment. One experiment was conducted for individual nurse and another for all nurses.

5.1 Results on individual nurse

We plotted the accuracy rate of each nurse in different cross validation settings for Random Forest, Binary Relevance and Random Forest SRC. In the plots x-axis shows the number of cross validation and y-axis shows accuracy in percent. The performance measures are also shown in separate plots for individual nurse in percent.

In Figure 5 the accuracies for each nurse are shown in different cross-validation settings. For Random Forest algorithm we can see that the accuracies are comparatively higher for 20-fold cross validation. For nurse 4 the accuracy rate is around 16% for 5-fold, the overall accuracy rate was around 14-15% for 20-fold. The reason is the number of instances for each nurse is not very large so when we are taking 15 or 20 fold the model is estimating over small sized test case which is not sufficient.The results and impact of cross validation indicate that for this setting multiclass method will not perform well. Even if we take larger amount of data, the overall accuracy is not probably going to increase.

The results of Binary Relevance method show a continuous increase with the increasing number of folds. It can be observed from Figure 5 that for 20-fold cross validation we are getting the highest accuracy for each nurse.However the reason is same as mentioned above that it also predicted over small set of data. With increase of data this method might perform better than Random Forest.

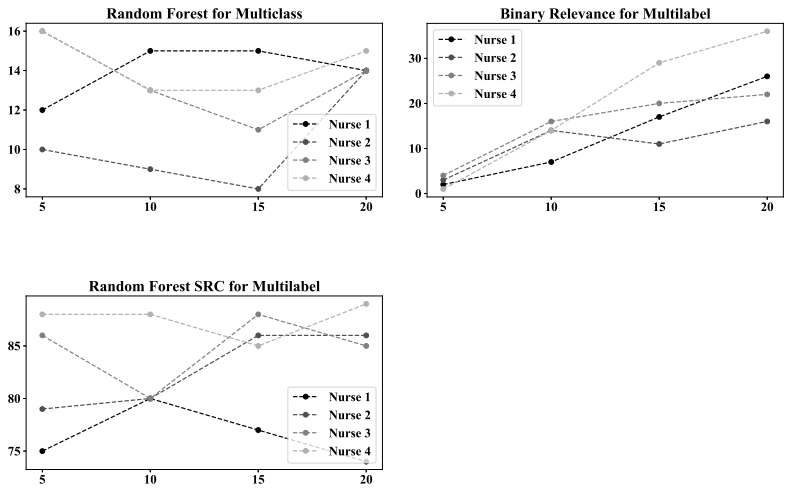

Figure 5: Accuracy on different cross validation

For Random Forest SRC, the accuracy rate over the set of cross validation is different. For nurse 1 10-fold is good, for nurse 2 and 3 15-fold is good and over 20-fold we got good result for nurse 4. The difference here from other two methods is irrespective of different cross validation the accuracy rate for is much higher.

In Figure 5 we calculated the percentage correct number of predictions over total samples for multiclass classification. We measured accuracy this way because only there is only one sample per instance so the correct number of predictions is enough to find the overall accuracy. For multilabel we calculated subset accuracy because we wanted to find the number of instances where each label was correctly predicted. The purpose of using cross validation over the data was to avoid over-fitting and observe the changes in accuracies. Figure 5 also shows each nurse having different accuracies for different folds. The reason is we are randomly shuffling our data and dividing it into different folds. In case of each nurse we have some activity classes that were performed very few times. If those activities are included in test fold but not in train fold, it will lead to lower accuracy (Nurse 2 15-fold for Random Forest). This is how the number of folds in cross validation is affecting the results for each nurse. For the cross validation experiment we only considered the tuning for k (the number of fold) as hyper-parameter in order to observe the change in accuracies and avoid the model from getting over-fitted.

In Figure 6, we can find that the performance measures of nurse 1 is comparatively higher for 10, 15 and 20-fold cross validation varying from 12-14%. For nurse 2, a 20-fold cross validation provides good result in between 10-14%, for nurse 3 and 4 5-fold is good both of their performance measures are varying from 13-15%. We calculated weighted precision, recall and F1 score. Because weighted calculates

metrics for each label, and find their average weighted by support (the number of true instances for each label). Weighted metrics take imbalanced data into account so in our problem settings these metrics should be measured. According to the findings we can say that for nurse 3 and 4, 5-fold cross-validation is good enough to get a good performance measure. For nurse 1 and 2 10-fold is suitable. Therefore, the changes in cross-validation is also changing the performance.

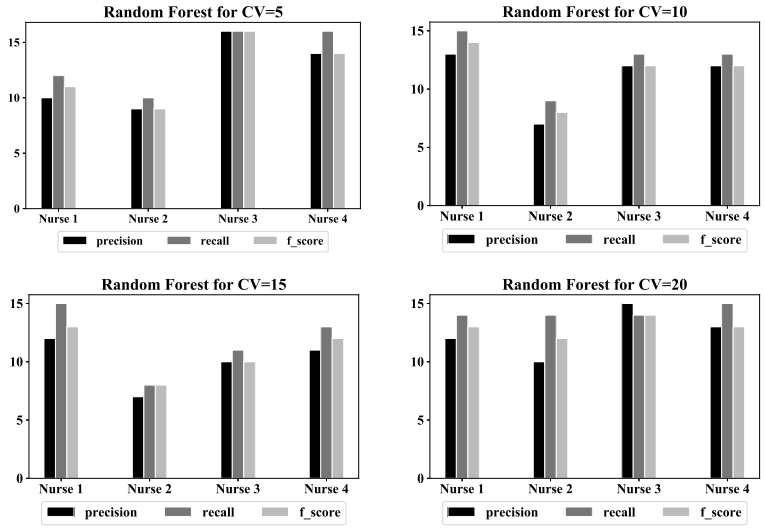

Figure 6: Performance measures of nurses with Random Forest

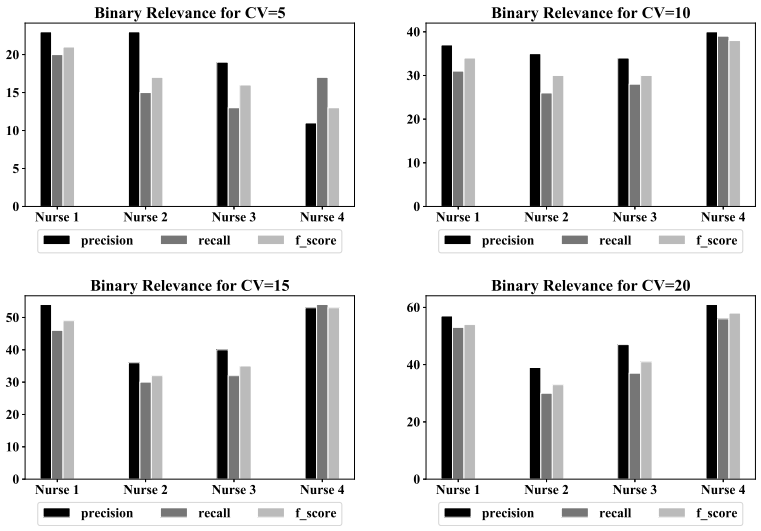

Figure 7: Performance measures of nurses with Binary Relevance

In Figure 7, after applying Binary Relevance we found that with the increasing number of folds in cross validation there is a bit increase in the performance measures. For nurse 1 the precision, recall and F1 score was varying from 10-15% and with increased number of folds the performance increased above 50%. For nurse 3 with 40-50% and nurse 4 with 50-60% a 20-fold cross validation is good enough. For nurse 2 the performance increased with 10-fold around 30-40% but did not change much for 15 or 20-fold cross validation. In order to get this result we also measured weighted precision, recall and F1 score; the reason is same as we mentioned earlier.

Figure 8 shows we achieved good results in case of Random Forest SRC for each nurse in 10, 15 and 20-fold cross validation setting and the results are quite stable. It means for smaller or larger test data it is going to perform well.The precision, recall and F1 score are varying between 60-80% for each nurse. For other two methods there is a little impact of cross validation. The fact is with multiple rounds of cross validation the estimated model's prediction should be better. Weighted precision, recall and F1 score is used here due to assigning the weight of each label as the same way we showed in Section 3.

Why Random Forest SRC performs well: RF differs from the RF SRC as it grows binary trees based on randomization procedure. It also uses another layer of randomization by using random feature selection. Rather than splitting a tree node using all features, RF selects at each node of each tree, a random subset of features that are used to split the node. The purpose of this two-step randomization is to de-correlate trees and reduce variance. However, Random forest SRC uses Gini index splitting rule [7]. Such splitting rules possess an end-cut preference (ECP) splitting property which is the property of favoring splits near the edge for noisy variables.

Thus it maximizes the tree node sample size and makes it possible for the tree to recover from the split downstream[22]. Random forest SRC that is using Gini index splitting rule possesses this useful property that makes the algorithm performing better.

For Random Forest the overall performance is not very high. In case of Binary Relevance it may not provide good result in accuracy but the results of other metrics are good. It means multiclass classification is not suitable for the prediction of our problem setting. Even though having impact of cross validation on Binary Relevance, we can see the performance is not higher than Random Forest SRC. Therefore, it is clear that multilabel classification methods are good enough to solve our problem setting.

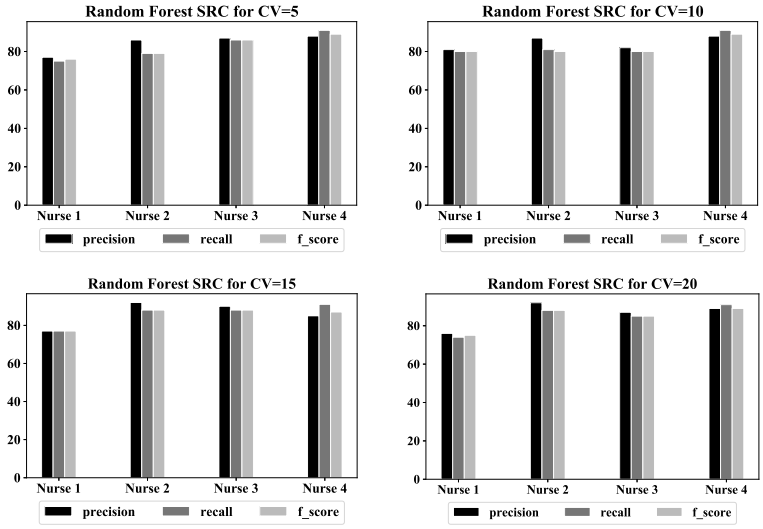

Figure 8: Performance measures of nurses with Random Forest SRC

5.2 Results on all nurses

We got some intuition after applying different set of cross validation for each nurse. However in order to confirm the performance over a larger size data we combined all nurses data. We applied three methods on the combined data. In this case, we train our data on Nurse 1, 2, 3 and test the data on Nurse 4. The reason for choosing such train test split is that; there are many activities that nurses do not have in common. In order to train the methods in a proper way we chose those three nurses

because most of their activities are similar with Nurse 4. For example, if nurse 1, 2 have performed [Record Observation Serving], nurse 3 has performed [Record Observation Wipe], nurse 4 has performed [Record Observation Serving, Wipe] as activities; we definitely need to consider nurse 1, 2 and 3 for training. Otherwise there will be so many activities unconsidered for training if they exist in test data.

In Figure 9, the performance of Random Forest SRC is better than the other two methods similarly as it was in case of individual nurse. We can see that for traditional Random Forest the accuracy, precision, recall and F1 score rate are in between 10-11%. Binary Relevance achieves 36% of accuracy; higher than Random Forest. The precision recall and F1 score are around 28%, 13% and 17% respectively. In this case. Random Forest SRC achieves much better accuracy than Random Forest and Binary Relevance do. The figure shows that accuracy about 66%. Precision, recall and F1 score are around 86%, 62% and 63% which is very promising result for this challenging dataset. In this plot subset accuracy is calculated for Binary Relevance and Random Forest SRC as they are multilabel methods and other three metrics are same for all three methods. For multiclass Random Forest the percentage of correct predictions was calculated.

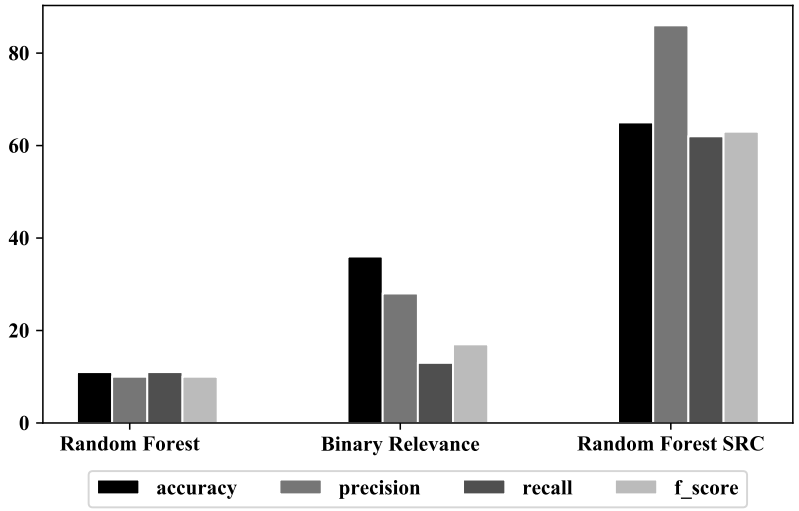

Figure 9: Performance of RF, BR and MLRF on all nurses data

6 Discussion and Motivation

In this paper we used a realistic nursing dataset to classify nursing activities by considering it as a multilabel classification problem. From the results we notice

that both in case of individual and combined nurse data the performance of three methods are similar. In order to find a thorough analysis on the data we used k-fold cross validations in different settings. In our case the role of cross validation important because how much data we are considering in larger or smaller number of folds gives us a better picture of the performances of methods we applied. Among them Random Forest SRC achieves highest accuracy of above 70% on average for each nurse. The performance drops a little in case of combined nurse data experiment. The reason is that the way of performing same activity by multiple person is also different. Another reason is we considered as much data with similar label as we could with an aim to train the models properly; the problem is there are still some activities that only exist in test data. So, predicting those activities is not possible with this experimental setting.

The most challenging part of this experiment was data preprocessing. While processing the data we were not able to find the location, pressure and temperature values of certain time. Even though we had the action labels for that particular time, we had to consider it as missing data and proceeded without them. Nurses performed some activities only for one or two times; we removed them as they were making our dataset highly imbalanced. At first, we were trying to run the experiments with normal statistical features like mean, standard deviation, variance etc. but most of the extracted features were not well correlated. So, we extracted some additional features such as velocity acceleration to achieve proper result.

We could achieve clearer picture about the whole data. The limitation of our work is we used small amount of data to conduct our experiment. For a better training and testing we need a sufficient amount of data for each subject. The action history contains one day data for each nurse. With insufficient data we might have difficulty in gaining more insightful information and good results. It is true that our dataset in smaller; however it is really very difficult to collect such data because it takes several days and huge amount of efforts. We had to collect user consent from nurses, one person was always assigned there 12 days and we had to pay for every day that becomes really costly process. Even though the dataset is smaller it is very much precious in a sense we collected activities related to hand hygiene, dataset contains multiple labels, focus on entry to exit data record and most importantly the labels are annotated by the observer. This type of dataset is rarely available in public. We used k-fold cross validation and increased the folds gradually to avoid over-fitting problem. Our assumption is that for both smaller or larger data with multilabel; exploiting label relevance can play a significant role.

As evaluation metrics, we considered four performance measures to evaluate our findings. We calculated the accuracy,weighted precision, recall and F1 score. Instead of taking micro or macro average we considered the weighted one because micro average calculates metrics globally by counting the total of true positives, false negatives and false positives. Macro calculates metrics for each label and find their unweighted mean. It does not take label imbalance into account. On the other hand, weighted calculates metrics for each label, and find their average weighted by support (the number of true instances for each label). Unlike macro it takes the imbalance label into account. The choice of evaluation metrics can be varied based

on the nature of dataset and experiment. In our case we were aware of the positive predicted value and true positive rate. So, we used precision, recall and F1 score.

Regarding this experiment, our intuition is that we can exploit label information based on the complete or partial relevance among labels. In such case we need to find those labels which are co-occurring frequently. It looks quite easier and efficient to exploit only those labels instead of checking one by one. Another assumption is that we can utilize neural network in order to resolve this multilabel problem. The benefits of using neural network is we do not need to extract additional features that can optimize our task. Besides, neural network can predict multiple output nodes for each instance. So, it is possible to predict the multiple labels based on the probabilities of output nodes.

After summarizing all our findings, we can conclude that activity recognition in a realistic scenario is one of the most challenging and different from traditional activity recognition. Because the nature of the data is quite different from data collected in controlled environment. In controlled environment activities are limited, duration is known and data is collected under a fully observed area whereas in real life nurses are performing multiple activities at a time and no fixed duration of any of them. So, it is highly possible that there could be multiple actions within a specific time window. Analyzing the realistic data is the biggest motivation behind this entire work.

7 Conclusion and Future Work

In this paper we presented how smartly we can utilize different methods in activity recognition. In our problem setting, we worked with a challenging realistic nursing dataset to recognize their activities. We applied both multiclass and multilabel methods to prove that for such condition traditional algorithms cannot solve the problem. Three classifiers are considered for recognition study: Random Forest and Binary Relevance and Random Forest SRC. Among them, Random Forest SRC achieved better result for individual and combine data settings. The reason behind the good performance is that the time window we considered as an instance actually was containing multiple labels that could not be predicted by multiclass algorithm. This type of nursing activity recognition is very much important especially in providing efficient healthcare, infection control and hand hygiene. Our work leads to many future directions such as exploiting label information, analyzing partial relevance. Our next step will be to exploit complete or partial relevance among labels and apply simple neural network in order to utilize this multilabel classification in activity recognition more efficiently.

References

1. Oscar L, Miguel L (2013) A Survey on Human Activity Recognition Using Wearable Sensors. Communications Surveys Tutorials, IEEE, vol,15, pp.1192-1209.
2. Sozo I, Naonori U, Yasunobu N, Naoki N (2015) Mobile Activity Recognition for A Whole Day: Recognizing Real Nursing Activities with Big Dataset. ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 1269-1280.
3. Sozo I, Naonori U, Yasunobu N, Naoki N (2016) Understanding Nursing Activities with Long-term Mobile Activity Recognition with Big Dataset. ISCIE International Symposium on Stochastic Systems Theory and Its Applications, pp. 1-11
4. Takebe Y, Kanai-Pak M, Kuwahara N, Maeda J, Hirata M, Kitajima Y, Ota J (2013) Recognition of Nursing Activity with Accelerometers and RFID. Kybernets, vol.42, pp.1059-1071
5. Momen K, Fernie GR (2010) Nursing Activity Recognition Using an Inexpensive Game Controller: An Application to Infection Control. Techno Health Care, vol.18, pp.393-408
6. Grigorios T, Ioannis K, (2009), Multi-label Classfication: An Overview. International Journal of Data Warehousing and Mining, vol.3, pp.1-13
7. Hemant I, and Udaya B K (2016) randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC)https://kogalur.github.io/randomForestSRC/theory.html
8. Liming C, Jesse H, Chris D N, Diane J C, Zhiwen Y (2012) Sensor Based Activity Recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 6, pp. 790-808
9. Hande O, Cem E (2010) Wireless Sensor Networks for Healthcare: A Survey. Computer Networks: The International Journal of Computer and Telecommunications Networking, vol. 54,pp.2688-2710
10. Ye L, Liqiang N, Li L, David S R (2016) From Action to Activity: Sensor-Based Activity Recognition. Neurocomputing, vol.181, pp.108-115
11. Alexandros P, Nikolaos G B (2010) A Survey on Wearable Sensor-Based Systems for Health Monitoring and Prognosis. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol.40, pp.1–12
12. Jennifer R K, W. Gary, M W, Samuel A M (2011) Activity Recognition Using Cellphone Accelerometers. ACM SIGKDD Explorations Newsletter, vol.12, pp.74-82.
13. Fernanno M R, Rene G, Gernot A F, Sascha F, Michael T H (2018) Convolutional Neural Networks for Human Activity Recognition Using Body-Worn Sensors. Informatics, vol.5.
14. Ravi N, Nikhil D, Preetham M, Michael L L (2005) Activity Recognition From Accelerometer Data. AAAI, vol.3, pp.1541-1546.
15. Rahul K, Imroj Q, Jaskaran SV, Narayanan CK (2015) Multi-label Learning for Activity Recognition. International Conference on Intelligent Environments, pp. 152-155
16. Ehsan M, Ricardo C, Fernaendo T, Mahmood F (2014) Mahmood, Multi-label Discriminative Weakly-Supervised Human Activity Recognition and Localization. Asian Conference on Computer Vision. doi: 10.1007/978-3-319-16814-2.16
17. Thuong N, Sunil G, Svetha V, Dinh P (2014) A Bayesian Noparametric Framework for Actvitiy Recognition Using Accelerometer Data. 22nd International Conference on Pattern Recognition, pp. 2017-2022
18. Raihani M, Muhammad Z, Sulaiman M N, Thinagaran P (2018) Multi-label Classification for Physical Activity Recognition from Various Accelerometer Sensor Positions. Journal of Information and Communication Technology, vol.17, pp.209-231
19. Alaa A, Vaidehi M, Doreen B, Ralf S (2016) Activity Recognition in Multi-User Environments Using Techniques of Multilabel Classiciaiton. 6th International Conference on Internet of Things pp. 15-23
20. Oresti B, Juan-Manuel G, Miguel D, Hector P, Ignacio R (2014) Widnow Size Impact in Human Activity Recogntion. Sensors, vol. 14, pp. 6474–6499.
21. Philipp P, Quay A, Giuseppe C, Clemens S, Bernd B (2017) Multilabel Classification with R Package mlr. R Journal, vol.9, pp. 352-369
22. Hemant I (2014) The Effect of Splitting on Random Forests. Machine Learning, 99. 10.1007/s10994-014-5451-2.