User Activities Outlier Detection System Using Principal Component Analysis and Fuzzy Rule-Based System

Sawsan M. Mahmoud ‡ Ahmad Lotfi † Caroline Langensiepen †

ABSTRACT
In this paper, a user activities outlier detection system is introduced. The proposed system is implemented in a smart home environment equipped with appropriate sensory devices. An activity outlier detection system consist of a two-stage integration of Principal Component Analysis (PCA) and Fuzzy Rule-Based System (FRBS). In the first stage, the Hamming distance is used to measure the distances between the activities. PCA is then applied to the distance measures to find two indices of Hotelling’s $T^2$ and Squared Prediction Error (SPE). In the second stage of the process, the calculated indices are provided as inputs to FRBSs to model them heuristically. They are used to identify the outliers and classify them. Three case studies are reported to demonstrate the effectiveness of the proposed system. The proposed system successfully identifies the outliers and helps in distinguishing between the normal and abnormal behaviour patterns of the Activities of Daily Living (ADL).

Categories and Subject Descriptors
J.4 [Computer Applications]: Social and Behavioural Sciences; D.2 [Data]: Data Storage Representations; H.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms
Human factors

‡ School of Science and Technology, Nottingham Trent University, Clifton Lane, NG11 8NS, Nottingham, United Kingdom. Email: sawsan.mahmoud@ntu.ac.uk
† School of Science and Technology, Nottingham Trent University, Clifton Lane, NG11 8NS, Nottingham, United Kingdom. Email: ahmad.lotfi@ntu.ac.uk
† School of Science and Technology, Nottingham Trent University, Clifton Lane, NG11 8NS, Nottingham, United Kingdom. Email: caroline.langensiepen@ntu.ac.uk

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

PETRA '12, June 6 - 8, 2012, Crete Island, Greece.
Copyright ©2012 ACM ISBN 978-1-4503-1300-1/12/06 ...$10.00.

Keywords
Intelligent Environments, Smart Homes, Outlier Detection, Abnormal Behaviour, Principal Component Analysis, Fuzzy Rule-Based System, Distance Measure

1. INTRODUCTION
It is very important to find a way identifying outliers or anomalies within data and show how they affect the data analysis. Outliers are those data that vary extremely from other data points. Outliers may cause problems in the performance of any system. Therefore, an outlier detection technique could help in increasing the overall performance of the data analysis [1, 2]. For example in a smart home environment, early detection of any anomalies in the behaviour of the occupant would lead to further investigation and understanding the behavioural pattern of the user.

In this paper, a two-stage user activities outlier detection system is introduced. The proposed system is implemented in a smart home environment equipped with appropriate sensory devices. The sensory devices produce long sequences of data as binary time series, indicating presence or absence of the occupant in different areas. The proposed system is used to identify outliers and then classify how far the outliers are from the remaining normal data.

In the first stage of the proposed system, the actual binary time series data is used to construct a distance matrix that indicates the degree of the differences between them. The distance matrix represent the distance between each individual observation. The outliers in multi-dimensional data do not appear by using each individual dimension i.e. they are not identified using a univariate approach (in this approach outliers are detected when the standardised value of the data point is large). Therefore, it is better to handle the outliers using a multivariate approach where outliers are detected considering all features of the multi-dimensional data [3]. An approach which efficiently reduces the high dimensionality of the data is required for better visualization and ultimately identification of any outlier or abnormality. Principal Component Analysis (PCA) is already being used for dimensionality reduction [4]. PCA transforms the actual multi feature data into new reduced features called “principal components” which are uncorrelated [5]. The principal components are used in the calculation of two indices including $T^2$ Hotelling and Squared Prediction Error (SPE) to be provided into the second stage of outlier detection system.

In the second stage, the outputs from PCA will be the inputs to a group of Fuzzy Rule-Based Systems (FRBSs). Information provided by the indices generated from the PCA
are fuzzified into fuzzy values. Then, the FRBS classify the features into fuzzy labels. FRBSs are used to deal with the uncertainty associated with human activities. The FRBS summarises a user’s activities. They are necessary to describe the trends of the occupant’s behavioural patterns and provide answers to queries related to any deviation from these patterns. In addition, using FRBS will help in understanding and simplifying addition, removal and modification within the knowledge structure [6]. A FRBS is constructed for each sensory device and the final decision is made based on a combination of the outputs from all individual FRBSs.

Three case studies are presented in this paper to illustrate how the proposed system will work.

This paper is organized as follows: Section 2 presents a summary of the related works on the identification of outliers in a smart home environment. In Section 3, our proposed outlier detection algorithm is introduced with a brief description of different components in the algorithm. The implementation of the proposed algorithm and its validation on three different case studies are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. RELATED WORKS

Many outlier detection techniques are proposed in statistics including mean, standard deviation and z-score (also known as standard score). For example, the authors in [2] investigated and built an outlier detection mechanism to improve energy management in a smart environment. The authors identified the outlier by finding the extremes using standard deviation. They also included a mechanism to rank the identified outliers to measure severity of the outlier. The z-score technique is used in [1] to detect outliers in ADL. The results are presented and tested using a volunteer in an apartment setting. As the data is not normally distributed, mean and standard deviation are not good measures for detecting outliers.

An alternative approach is the box plot which is a graphical representation approach for examining data sets. For example, box plot is used in [7] to identify outliers which lie unusually far from the main body of the data. A box plot displays five important data summaries. They are: lowest value, lower quartile, median, upper quartile, and highest value. A solid line drawn across the box between the upper and lower quartiles to locate the median. The lowest and highest values exist at the boundary of the solid line. The advantages of the box plot are that it can display differences between populations without making assumptions about the underlying statistical distribution and the distance between the parts of the box indicates the degree of spread and skewness in the data set. However, it is argued that box plot is not an appropriate approach for every kind of data.

A one class Support Vector Data Description (SVDD) algorithm is used in [8] to distinguish between normal and abnormal behavioural patterns. Using this algorithm, the abnormality is detected from outliers of the class and a boundary around the target data is made by enclosing the target data within a minimum sphere. Several infra-red motion sensors are used to monitor an elderly person living in a smart environment to improve their personal healthcare system.

Generally, finding outliers in a high dimensional vector is a complex process. There are several methods which are used to find outliers in low dimensional space including Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD)[3, 9, 10]. To process data with a high dimension, PCA has proven to be the preferred option. For example, in [10], an outlier detection system is introduced using PCA technique in a combination with a hierarchical clustering technique. They used cluster principal component analysis as a new distance-based method. The system is able to identify outliers in both single and multi-dimensional data.

On the other hand, fuzzy systems are widely used in smart homes to detect events in a wireless sensor network [10, 11, 12, 13]. For examples in [14], the authors proposed an algorithm using FRBS to deal with uncertain information. In addition, a system for ADL recognition system is proposed in [12] using fuzzy system. Linguistic terms are used to monitor the ADLs of an occupant to offer him/her a safe, more comfortable and appropriate environment. However, in that research the fuzzy logic has been applied only on the simulated data. In contrast with reported research in this area, the proposed outliers detection system in this paper is implemented using PCA in a combination with FRBS for abnormality and outlier detection in ADLs.

3. OUTLIER DETECTION SYSTEM

Most data sets collected from real environments contains some unusual observations. Observations under the same conditions with extreme values compared with the other observations are considered as outliers. Outlier data points are either different from other observations in data points or they are not under the same probability model of the data. There are different kinds of outliers such as data errors, unusual values or events etc. [1, 2, 3]. This paper presents a two stage outlier or anomalies detection system for ADLs in a smart home. The first stage includes a dimensionality reduction of data in which PCA is used. The second stage includes an outlier or abnormality identification using FRBS. The architecture of the proposed outlier detection system is depicted in Figure 1. Components of the proposed system to identify the outlier are listed below:

- Collect the sensor data from the environment.
- Calculate the distance matrices for the collected data using distance measure.
- Conduct the PCA for the distance matrices. The output from the PCA are Hotelling’s \( T^2 \) and SPE indices.
- Calculate confidence limits for both Hotelling’s \( T^2 \) and SPE indices. This will be identified as the universe of discourse for the FRBSs.
- Compute the degree of memberships for both Hotelling’s \( T^2 \) and SPE indices as the inputs of FRBSs.
- Formulate the rules in the FRBSs and check whether inputs exceed the confidence limits.
- Defuzzify the output values to provide a rank for data point to indicate the membership degree of the outlier.

When new data are available, any outlier or anomalies are detected and ranked accordingly. More details about the components described above are presented in the following sections.
3.1 Distance Matrix

Distance or dissimilarity measures are used to find the degree of differences between two sets of data. Generally, finding dissimilarities between two vectors mean that both vectors have changes in patterns of attributes. The values that have common patterns are considered to be closer to one another than those with different patterns [15, 16]. Without any loss of generality, in this paper only binary distance or dissimilarity measures are investigated. Most low sensory data collected from a real environment are based on occupancy sensors presenting in long sequence of binary data. There are many distance measures for binary vectors suggested in the literature [15, 17, 18, 19, 20]. The most commonly used distance measure between two binary vectors is Hamming distance [21]. Hamming distance can be defined as the number of mismatching bits between two binary vectors of the same length. Formally, let \( A \) and \( B \) be two binary vectors i.e. have the \( i^{th} \) feature value either 0 or 1. \( A \) and \( B \) are representing two sets of binary data collected for two separate periods of time for the same sensor. The Hamming distance \( D \) between \( A \) and \( B \) is computed as follows:

\[
D(A, B) = S_{01} + S_{10}
\]

where \( S_{01} \) is the number of occurrences of matches with 0 in the first pattern and 1 in the second pattern at the corresponding positions. \( S_{10} \) is the number of occurrences of matches with 1 in the first pattern and 0 in the second pattern at the corresponding positions. Readers are referred to [22] for more details where Hamming distance measure is used to monitor the behaviour patterns in an intelligent environment.

3.2 Principal Component Analysis

Principal component analysis is a statistical tool commonly used to reduce the dimensionality of a data set consisting a large number of interrelated variables, while preserving the variation within the data set [4]. PCA is a technique which helps in producing a better visualization of the data since it transform the data in a way that shows the maximum variability within the data. In general, PCA can be used to understand the main features of the data, reducing the dimensionality of multi-dimensional data, finding the relationships between variables of the data, and identifying the trends in the data [23, 24]. PCA has been used in many real-time processes monitoring for detecting changes in operating points, sensor faults, process faults, and plant disturbance [25].

Formally, let us consider \( X \) as an input matrix which consists of \( N \) observations and \( M \) variables. Standardization such as z-score is normally used when variables are measured in different units. PCA takes the input matrix and transforms it into two eigenvectors, \( e_1, e_2, \ldots, e_k \), and associated eigenvalues, \( \lambda_1, \lambda_2, \ldots, \lambda_k \), where \( k \) is the number of selected PCs. The eigenvectors contain the principal components (PCs). The first PC contains the data with the highest variance while the second PC contain the data with the next highest variance and so on for other components. The format of the eigenvalues and eigenvectors are as follows:

\[
\Lambda = \begin{bmatrix}
\lambda_1 & 0 & \ldots & 0 \\
0 & \lambda_2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \lambda_M \\
\end{bmatrix}
\]

\[
V = [e_1, e_2, \ldots, e_k]
\]

Changes within the data are detected using two statistical index measures named Hotelling’s \( T^2 \) and Square prediction error (SPE). When PCs are identified, the \( T^2 \) and SPE indices measures are computed. These two statistic measures are briefly described below.

3.2.1 Hotelling’s \( T^2 \) Statistic
Hotelling’s $T^2$ measures the squared norm of the current sample from the centre of the normal data points region [25]. In other words, the $T^2$ index measures the variations in the PCs and it is calculated using the following expression:

\[ T^2 = X^T V \Lambda^{-1} V^T X \]

or

\[ T^2 = \sum_{i=1}^{k} \frac{t_i^2}{\lambda_i} \]

where $t_i$ is the $i^{th}$ element in the vector $t = V^T X$. The limit of $T^2$ index with a confidence level $\alpha$ is:

\[ T_{lim} = k(N-k) \] \[ \frac{F(k,N-k,\alpha)}{N-k} \]

where the $F(k,N-k,\alpha)$ corresponds to the probability point on the F-distribution with $(k,N-k)$ degrees of freedom and confidence level $\alpha$.

3.2.2 Square Prediction Error Statistic

The Square Prediction Error (SPE) index measures the projection of the data points on the residual subspace [25]. It is calculated using the following expression:

\[ r = X^T - X = X^T - V_k V_k^T X \]

and

\[ SPE = r^T r \]

The residual matrix $r$ captures the variations in the observation space spanned by the PCs associated with the $M$ smallest singular values. The sub spaces spanned by $X$ and $X^T$ are called the score space and residual space respectively [9]. The limit for SPE index which denotes the upper control limit for SPE index with a confidence level $\alpha = 95\%$ is:

\[ SPE_{lim} = \theta_1 \left[ \frac{C_{h_0} \sqrt{2h_2 \theta_1}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{\frac{1}{2}} \]

where $\theta_1 = \sum_{j=k+1}^m \lambda_j$ and $h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2}$.

The above two indices are checked whether they exceed their control limits. If both $T^2$ and SPE measures exceed their control limits then the status for the process is abnormal. Otherwise, the status of the process is considered as normal when both indices are less than their limits [25]. It is also possible to reach other options if none of the above are satisfied. Therefore, to reach a decision based on the values of the indices, a fuzzy rule-based system is used to provide the decision.

3.3 Fuzzy Rule-Based System

A fuzzy rule-based system can efficiently model the vague or uncertain data in sensor networks. A FRBS can report the user activities in terms of linguistic variables instead of the raw or pre-processed sensor data. In addition, it can give justifications for that report to be more understandable to the end user of the system [11, 12, 13, 26]. FRBS are used in many applications including: automated diagnosis, control systems, image processing and pattern recognition. Fuzzy if-then rules of the following configuration are employed for the modelling of the PCA statistical measures to identify the outliers:

\[ R_i^j: \text{If } T_{2j}^i \text{ is } \tilde{A}_i^j \text{ and } SPE_j \text{ is } \tilde{B}_j^i \text{ then outlier } \text{rank}_j \text{ is } \tilde{C}_j^i \]
where $R_j^i$ is the label of $i^{th}$ rule for the sensor $j$. $T_{lim}^j$ and $SPE_j$ are the inputs for the sensor $j$. $\text{outlier}_{rank_j}$ is the output. $A_j^i$ and $B_j^i$ ($i = 1, 2, ..., m$ and $j = 1, 2, ..., p$) are fuzzy labels (fuzzy values) for inputs and $C_j^n$ ($i = 1, 2, ..., p$) is the label for outputs. $p$ is the number of sensor data set, $m$ is the number of labels for input membership functions and $n$ is the number of labels for output membership functions. The universe of input variables is limited to $T_{lim}$ and $SPE_{Elim}$ as calculated in the previous section. The final outlier rank is decided based on the rank of the outlier for each sensor. The final rank is calculated as:

$$\text{outlier}_{rank} = \min_{j=1}^p (\text{outlier}_{rank_j})$$

### 4. IMPLEMENTATION

In this section, the proposed outlier detection system is validated using three different case studies. They are based on data collected from three different real environments. In each environment, there is a unit which receives data from the sensors and uploads it to a central database. In this study, only door entry and occupancy sensors are used. Collected data is recorded with a resolution of one second and is based on a single occupant environment. More detailed analysis of these case studies is presented below.

#### 4.1 Case Study I

The data for this case study is collected from a house equipped with JustChecking monitoring system [27]. The data is collected for a period of one year for an elderly occupant living alone. For this case study, two door entry sensors including front door and back door as well as four motion sensors including kitchen, bedroom, bathroom and lounge are used. In the beginning of our analysis, the binary data is collected for a period of one year for an elderly occupant living alone. For this case study, two door entry sensors including front door and back door as well as four motion sensors including kitchen, bedroom, bathroom and lounge are used. In the beginning of our analysis, the binary data is recorded with a resolution of one second and is based on a single occupant environment. More detailed analysis of these case studies is presented below.

The data for this case study is collected from a house equipped with JustChecking monitoring system [27]. The data is collected for a period of one year for an elderly occupant living alone. For this case study, two door entry sensors including front door and back door as well as four motion sensors including kitchen, bedroom, bathroom and lounge are used. In the beginning of our analysis, the binary data is recorded with a resolution of one second and is based on a single occupant environment. More detailed analysis of these case studies is presented below.

For each FRBS, there are nine rules with three membership for each input and five membership for each output ($m = 3$ and $n = 5$). Fuzzy Labels for inputs are defined as high, medium and low. For the output of the FRBS representing outlier rank, the fuzzy labels are: Extreme Outlier (EO), Slight Outlier (SO), Medium (M), More or Less Normal (MN) and Normal (NO). The confidence limits for both SPE and Hotelling’s $T^2$ measures are computed for each sensor data set. For example, in this case study, the SPE limits are 1.9255, 6.2451, 211.0567 and 20.9023 for back door, front door, lounge and kitchen sensor respectively. In these figures, the text numbers are the days that activities are carried out. In these figures, the PCs are calculated using $\alpha = 95\%$. PCs are used to calculate two indices namely residual SPE and Hotelling’s $T^2$ measures. The SPE and Hotelling’s $T^2$ measures for the back door sensor is shown in Figure 4. These measures are used in the second stage of the process to classify outliers within the data sets. These indices are the inputs to FRBS classifiers with one output representing the outlier rank.

#### Figures

- **Figure 1**: The scree plot is shown in Figure 2 based on the kitchen motion sensor data. The first two PCs were selected. Figure 3-a to Figure 3-d show the scatter plots for the first and the second PCs of the binary data collected from the environment for back door, front door, lounge motion and kitchen motion sensors respectively. In these figures, the text numbers are the days that activities are carried out. In these figures, the PCs are calculated using $\alpha = 95\%$. PCs are used to calculate two indices namely residual SPE and Hotelling’s $T^2$ measures. The SPE and Hotelling’s $T^2$ measures for the back door sensor is shown in Figure 4. These measures are used in the second stage of the process to classify outliers within the data sets. These indices are the inputs to FRBS classifiers with one output representing the outlier rank.

For each FRBS, there are nine rules with three membership for each input and five membership for each output ($m = 3$ and $n = 5$). Fuzzy Labels for inputs are defined as high, medium and low. For the output of the FRBS representing outlier rank, the fuzzy labels are: Extreme Outlier (EO), Slight Outlier (SO), Medium (M), More or Less Normal (MN) and Normal (NO). The confidence limits for both SPE and Hotelling’s $T^2$ measures are computed for each sensor data set. For example, in this case study, the SPE limits are 1.9255, 6.2451, 211.0567 and 20.9023 for back door, front door, lounge and kitchen sensor respectively. In these figures, the text numbers are the days that activities are carried out. In these figures, the PCs are calculated using $\alpha = 95\%$. PCs are used to calculate two indices namely residual SPE and Hotelling’s $T^2$ measures. The SPE and Hotelling’s $T^2$ measures for the back door sensor is shown in Figure 4. These measures are used in the second stage of the process to classify outliers within the data sets. These indices are the inputs to FRBS classifiers with one output representing the outlier rank.

For each FRBS, there are nine rules with three membership for each input and five membership for each output ($m = 3$ and $n = 5$). Fuzzy Labels for inputs are defined as high, medium and low. For the output of the FRBS representing outlier rank, the fuzzy labels are: Extreme Outlier (EO), Slight Outlier (SO), Medium (M), More or Less Normal (MN) and Normal (NO). The confidence limits for both SPE and Hotelling’s $T^2$ measures are computed for each sensor data set. For example, in this case study, the SPE limits are 1.9255, 6.2451, 211.0567 and 20.9023 for back door, front door, lounge and kitchen sensor respectively. In these figures, the text numbers are the days that activities are carried out. In these figures, the PCs are calculated using $\alpha = 95\%$. PCs are used to calculate two indices namely residual SPE and Hotelling’s $T^2$ measures. The SPE and Hotelling’s $T^2$ measures for the back door sensor is shown in Figure 4. These measures are used in the second stage of the process to classify outliers within the data sets. These indices are the inputs to FRBS classifiers with one output representing the outlier rank.

For each FRBS, there are nine rules with three membership for each input and five membership for each output ($m = 3$ and $n = 5$). Fuzzy Labels for inputs are defined as high, medium and low. For the output of the FRBS representing outlier rank, the fuzzy labels are: Extreme Outlier (EO), Slight Outlier (SO), Medium (M), More or Less Normal (MN) and Normal (NO). The confidence limits for both SPE and Hotelling’s $T^2$ measures are computed for each sensor data set. For example, in this case study, the SPE limits are 1.9255, 6.2451, 211.0567 and 20.9023 for back door, front door, lounge and kitchen sensor respectively. In these figures, the text numbers are the days that activities are carried out. In these figures, the PCs are calculated using $\alpha = 95\%$. PCs are used to calculate two indices namely residual SPE and Hotelling’s $T^2$ measures. The SPE and Hotelling’s $T^2$ measures for the back door sensor is shown in Figure 4. These measures are used in the second stage of the process to classify outliers within the data sets. These indices are the inputs to FRBS classifiers with one output representing the outlier rank.
Figure 5: Membership labels for input and output variables for the back door sensor.

Figure 6-a to Figure 6-d labelled PCs based on the fuzzy outlier rank are shown for the same results shown earlier in Figure 3-a to Figure 3-d. It is obvious that the outliers are detected according to their position in the actual binary data sets. In addition, the most frequent observations (normal group of data) were selected from similar ADLs carried out during the year. However, a few observations, which are located far from the normal group of data, represent set of an individual days where ADLs are unlikely to be carried out (abnormal group). For example, three clear groups of data are recognized in Figure 6-a, these are: (1) Normal group (in fuzzy label NO) represents set of the days in the year where the back door is not often opened and closed, (2) Abnormal or outlier group (in fuzzy label EO), which is located to the left of the graph (i.e. day index 43), represents the data where the back door remained open for a long time, and (3) More or less normal group (in fuzzy label MN) represents the set of the days where the door is frequently opened but for a short time. It can be concluded that the proposed outliers system gives a visual grouping to the data of similar features and detects outliers and anomalies.

Table 1: Table of fuzzy rule for outlier rank identification (Extreme Outlier(EO), Slight Outlier(SO), Medium(M), More or Less Normal(MN) and Normal(NO)).

<table>
<thead>
<tr>
<th>IF</th>
<th>THEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPE low</td>
<td>NO</td>
</tr>
<tr>
<td>medium low</td>
<td>MN</td>
</tr>
<tr>
<td>high low</td>
<td>M</td>
</tr>
<tr>
<td>medium medium</td>
<td>MN</td>
</tr>
<tr>
<td>high medium</td>
<td>MN</td>
</tr>
<tr>
<td>low high</td>
<td>SO</td>
</tr>
<tr>
<td>medium high</td>
<td>SO</td>
</tr>
<tr>
<td>high high</td>
<td>EO</td>
</tr>
</tbody>
</table>

4.2 Case Study II

The data for this case study is collected from an elderly occupant living in her apartment based on the system developed by [28]. Four motions sensors covering the lounge, kitchen, bedroom and corridor are used. Additionally, two door entry sensors were used to monitor the bathroom and the main entrance doors. If there are slight outliers and changes or extreme outliers could be sent to the carer to help the elderly. Figure 7 shows the scatter plots of the labelled PCs based on the fuzzy outlier rank for the bedroom motion sensor and bathroom door sensor data sets respectively. As shown in this figure, no extreme outliers are found in these data sets, there are only slight outliers (stars) and more or less change in her behaviour (circles). Therefore, in these cases an alarm should be send to the carer to check if there is something wrong.

4.3 Case Study III

The data for this case study is also collected from another environment with an elderly occupant using the JustChecking monitoring system. The elderly occupant was first prescribed some medications, and her health status got worse. Consequently, she was roaming around during the early hours of the day, and her behaviour was considered as abnormal. Then her first medications were replaced by new medications and she got better. In this work, the data collected from this environment are split into two separate groups. One group represents the data when she was not well and the second group when she got better.

To distinguish between these two groups, as with previous experiments, the combined PCA and FRBS are used. Figure 8-a shows the scatter plots for the first and the second PCs of the binary data collected from the environment for the front door entry sensor. The corresponding labelled PCs based on the fuzzy outlier rank are shown in Figure 8-b. There is a significant difference between these two groups. The first group shows the normal data (circles) since their SPE and Hotelling’s $T^2$ are less than the confidence limits, while the second group shows the extreme outliers or anomalies (squares) as their SPE and Hotelling’s $T^2$ are greater than the confidence limits. The results obtained from this case study has shown that proposed outlier detection system can effectively distinguish between the normal and abnormal data.

5. CONCLUSIONS

In this paper, the integration of principal component analysis and fuzzy rule-based system are investigated to identify outliers and anomalies in an intelligent environment. The intention was to determine the effect of both the residual SPE and Hotelling’s $T^2$ in improving the results of PCA and how they are utilized in the fuzzification process of a fuzzy rule-based system. It has found that the proposed system successfully distinguishes between the normal and abnormal or outliers data points. Three case studies are used to show the effectiveness of our detection system.

6. REFERENCES

Figure 6: Scattered plot for the 1st and 2nd principal components of the data used in case study I with classification. Triangles represent normal, squares represent extreme outliers, stars represent slight outliers and circles represent more or less normal pattern.


Figure 7: Scattered plot for the 1st and 2nd principal components of the data used in case study II with classification. Triangles represent normal, squares represent extreme outliers, stars represent slight outliers and circles represent more or less normal pattern.

Figure 8: Scattered plot for the 1st and 2nd principal components of the front door data used in case study III. a) without classification b) with classification.

activities using dissimilarity measurements. In 4th Int. Conf. on PErvasive Technologies Related to Assistive Environments, PETRA’11, 2011.

[23] W. Müller, T. Nocke, and H. Schumann. Enhancing the visualization process with principal component analysis to support the exploration of trends. Conf. in Research and Practice in Information Technology (CRPIT), Tokyo, Japan, 60, February 2006.


