Industrial Management & Data Systems

Critical analysis of smart environment sensor data behavior pattern based on sequential data mining techniques

Gebeyehu Belay Gebremeskel Chai Yi Chengliang Wang Zhongshi He

Critical analysis of smart environment sensor data behavior pattern based on sequential data mining techniques

Gebeyehu Belay Gebremeskel and Chai Yi
College of Automation, Chongqing University, Chongqing, China, and
Chengliang Wang and Zhongshi He
College of Computer Science, Chongqing University, Chongqing, China

Received 29 December 2014
Revised 25 March 2015
6 May 2015
18 May 2015
Accepted 18 May 2015

Abstract

Purpose – Behavioral pattern mining for intelligent system such as SmEs sensor data are vitally important in many applications and performance optimizations. Sensor pattern mining (SPM) is also dynamic and a hot research issue to pervasive and ubiquitous of smart technologies toward improving human life. However, in large-scale sensor data, exploring and mining pattern, which leads to detect the abnormal behavior is challenging. The paper aims to discuss these issues.

Design/methodology/approach – Sensor data are complex and multivariate, for example, which data captured by the sensors, how it is precise, what properties are recorded or measured, are important research issues. Therefore, the method, the authors proposed Sequential Data Mining (SDM) approach to explore pattern behaviors toward detecting abnormal patterns for smart space fault diagnosis and performance optimization in the intelligent world. Sensor data types, modeling, descriptions and SPM techniques are discussed in depth using real sensor data sets.

Findings – The outcome of the paper is measured as introducing a novel idea how SDM technique's scale-up to sensor data pattern mining. In the paper, the approach and technicality of the sensor data pattern analyzed, and finally the pattern behaviors detected or segmented as normal and abnormal patterns.

Originality/value – The paper is focussed on sensor data behavioral patterns for fault diagnosis and performance optimizations. It is other ways of knowledge extraction from the anomaly of sensor data (observation records), which is pertinent to adopt in many intelligent systems applications, including safety and security, efficiency, and other advantages as the consideration of the real-world problems.

Keywords Activity description, Behavioral pattern, Data mining, Intelligent systems, Smart space, Trajectories

Paper type Research paper

I. Introduction

Critical analysis of Smart Environment (SE) sensor data is a technological exploration of intelligent system's functionality and performance (Brian and C.J.H., 2007). Sensor technologies are related to human cognitive capture and visualize behavioral patterns, which adopted in almost every modern intelligent system. Personal (smart home), safety (traffic management, military security), healthcare (cognitive behavior), business (sales track) are few domains. Furthermore, industries (architectural control), environmental monitoring, and location-aware services are an additional potential area of smart technology applications. In these applications, sensors captured various properties of

The authors are very thankful to the anonymous reviewers for their useful comments. The works is supported by the National Natural Science Foundation of China under Grant No. 61004112 and the Third Stage Building of "211 Project" under Grant No. S-10218 and the project of Innovative Team Building under the Grant No. 2009091011.

Industrial Management & Data Systems
Vol. 115 No. 6, 2015
pp. 1151-1178

0263-5577
DOI 10.1108/IMDS-12-2014-0386

physical phenomena become a huge sensor data, which is challenging to explore and mine behavioral patterns in a traditional approach. The problem is more noticeable to analysis the hidden data relationship and searches useful information (Tao *et al.*, 2009; Joëlle *et al.*, 2005).

Behavioral Pattern Mining (BPM) is the process of segmenting the signal or sensors' paths, which are an essential to reveal, object trajectories and characteristics for fault detections. The approach is an investigation of smart space applications in relation to sensor technologies' performances. Sensor data captured from diverse sensor devices and annotated as the objects' situations or moments. Therefore, data schemas, precision or accuracy, units of measurement factors are a seductive issue of BPM (Jae-Gil *et al.*, 2007). However, the behavioral pattern in large-scale sensor data is challenging. Since the data do not have common features, and the process of data synthesizing is complex as a set of instructions and event patterns (Elad, 2004). The issues can be summarized as: as sensor technologies, and applications are increasing, sensor data handling and analyzing become a challenge; how sensor data behavioral patterns are developed by considering the domain contexts and actuators' behaviors towards abnormal behavior pattern detections and synthesizing?; sensor data are not completely stored, which involves real-time activities. How to combine and analyzes trend and active sensor data?; and what analytic tool is more capable to explore sensor data as its types (volume, velocity, variety, etc.), complexity, sensitive, and massiveness and other behavioral factors?

In this paper, we proposed Sequential Data Mining (SDM) technique for a critical analysis of sensor data behavioral pattern towards Abnormal Behavior Pattern (ABP) detection. The approach is an arrangement of instances or events in a sequential manner or standard to define the objects' trajectories behaviors. It is the process of exploring the data model, which design in a given sequence of sensor data to define objects' properties how and why deviate from its normal conditions. Moreover, an intelligent system is the agglomeration of advanced IT and smart technologies, which is pertinent to the dynamic applications of SDM. The technique of SDM for sensor design, pervasive computing and pattern mining (Parisa and Diane, 2011) is a systematic approach demystify the process. It is essential to address each sequence of the behavioral pattern, which support to describe the nature of sensor data (Yan *et al.*, 2008; Thomas *et al.*, 2006).

The ultimate goal of this paper is critically analyzing sensor data toward behavioral pattern mining using SDM techniques. Therefore, its contributions can be summarized as: introducing a novel idea and techniques to demystifying sensor data behavioral patterns as the context of the domain; optimizing DM applications as to explore sequential sensor events. SDM adapts to analysis sequence event patterns based on clustering and distance mining techniques for sizing clustered elements, then extend it into normal and abnormal sequence pattern characterizations. Proposed and present a generic distance-based clustering algorithms to define sensor data pattern behavior, which is significant to segment the boundary of any trajectories as normal and abnormal properties. The approach is also scalable to overcome challenges that happened in the process of SE performance optimizations and fault diagnosis.

The rest of the paper is organized into five sections. Section II summarizes the related and synonymous research works in the field of pattern mining and sensor technologies. In Section III, we discussed the sensor data features, models and model descriptions, pattern clustering and defining rules and algorithms. Section IV is the empirical analysis of the methods and the detail discussed. The last, Section V is the conclusion of this work, in which followed by the acknowledgments and reference for cited articles in this work.

II. Related work

Many research works in SE or intelligent environments enhancing to the advancement of sensor technologies, which focussed on its applications. The approaches are focussing on researchers endeavor and real supervision of the artifacts and certain assumptions of sensor performances for the smart space. However, the methods are not capable to address the ever growing sensor data handle, store and process challenges. Since the sensor data behavioral pattern developments are demanding more time, efforts, data labeling, and descriptions (Ramakrishnan and Rakesh, 2013; Dong *et al.*, 2005).

SDM-based pattern mining is a systematic approach to implement both supervised and unsupervised methods to scale up data patterns in normal and abnormal behavior (Parisa and Diane, 2010). The significant of using these combined methods are for two major reasons: the supervised method is essential to considering the knowledge of the existing system norms that made an effort on unsupervised approach gain better results; and the unsupervised method provides access to extract the unknown and hidden information or patterns in the large-scale sensor data (Tˆam and Bernt, 2006). The methods give a clear understanding for data characterizing, annotating and modeling. The process is pertinent to know how to define the patterns and why the abnormal behaviors are significant to the newly merged pattern. It also provides a tremendous application. The ABP detection is, therefore, a novel concept to investigate the cause of the problems or system faults (Piciarelli and Foresti, 2006). It is uncommon or unique trajectories behaviors (Wiliem *et al.*, 2008), pattern irregularity (Oren and Michal, 2005) of sensor data.

Sensor data are signals or motions, which are the intelligent agents perceives the state of the physical objects moment records of the sensor devices. It needs to analyze the applications and functionality of smart spaces to optimize its performances and securities. In SE, the activities are predefined continual tasks. However, the data pattern could vary as the object trajectories, location, and time. In certain conditions, such as the electric power-off, the tasks may be dis-continual, which causes the process incomplete or abnormal (Parisa *et al.*, 2011). Therefore, such complex and large-scale sensor data demand an advanced analytical tool to analyze the behavioral patterns, which tracked in each path span of time. SDM-based sensor data BPM is a generic and inference approach for sensor data analysis (Charu, 2013).

III. Significant measure of sensor data behavioral pattern mining

Sensor readings are not completely random that correlate to time and space or location on a given sensor record, which captures by sensor devices to analyze the events' behavioral patterns. For instance, the temperature measures in a room can be the heat or cool conditions records, which are two events (or parameters). Therefore, measuring such events patterns are explored in relation to sensor device locations and data capturing-time that shows the temperature turning point, which help to define the intelligent system performance and functionality in the room. Therefore, the significant measure of sensor data is the critical analysis of event sequence patterns as the magnitude of an objects trajectories or behaviors at the specified location and time (Ahmed and Nada, 2013).

Behavioral pattern mining is a complex and dynamic process, which is challenging for statistical or traditional approaches. SDM technique is a solution to demystifying sensor data handle and analysis based on trajectories toward abnormal behavioral pattern detection. Smart home acquires and prior sensors knowledge is essential to evaluate intelligent system physical sets in a way that can optimize various intelligent tasks for different goals, including safety and security (Yiu and Yau, 2006). The technology is advancing to big shopping malls, robotic intelligent environment, interactive conference and others social crowd places. The system supports the integration of heterogeneous devices that facilitate collaborative intelligent activities (Erfaneh *et al.*, 2012). In the past, for a significant measure of intelligent behaviors, various computational techniques have been proposed (Parisa and Diane, 2011). Many researchers detected the uncommon and unique behavior of the trajectories based on video-camera surveillance common behavior contexts. However, these approaches are applied to specific issues. On the other hand, sensor technology is dynamic and crosscutting fields, which need generic analytic approaches to infer and scale-up for various problems solving process. Moreover: sensor records or captured images are the unsupervised streaming data type, which need advanced tools to explore the implicit knowledge; SDM technique is capable to scale up both supervised and unsupervised methods for sequential streaming data as a fundamental and adaptive research in relation to fault diagnosis and smart space optimizations; and abnormal behavioral based knowledge discovery is an interesting task to measure intelligent system performance and detect faults.

3.1 Sensor data in an intelligent ecosystem

A sense is describing in various ecosystems, including electronic sensing, human sensing and others (Mu-Yen and Edwin, 2013). All sensor data are characterized by sequential and dynamic types. The technological (electronic and nontraditional) sensor data recorded and collected in intelligent or smart environments, which are interesting to exploit sensor data behavioral patterns toward fault or abnormal pattern detections. SDM is needed to explore the behavior and patterns of the sensor data to extract valuable and implicit knowledge. The techniques/algorithms of clustering and the nearest neighbor used to classify sensor data behavior (Bhatia and Deepika, 2013). It is the systematic way to solve SE challenges and faults. Sensor data hidden interactions and similarity measures are analyzing using various techniques, which includes the Hidden Markova Model (HMM) (Karthika and Sumathi, 2012) and Emerging Patterns Mining (Tao *et al.*, 2009).

Sensor data sequences are often linear or multi-dimensional arrangements with different lengths. As the application of intelligent systems growing at a dynamic pace, pattern mining based on one-dimensional (linear) trajectory and similarity of a given time series is not always an effective approach. For example, sensor data of similar movement patterns might appear in two or more sensor devices, which are different sampling rates of tracking and sensor devices combined with various speeds of the moving objects. Such multi-dimensional data sets can be explored by applying SDM-based similarity function. Let the objects are points that move in a two-dimensional plane (x, y) at a time t. Therefore, the sensors $S=[(s_1, t_1), (s_2, t_2), \ldots, (s_n, t_n)]$ or s_i are the pair of (s_i, x, s_i, y) with their corresponding time t_i and (s_i, t_i) are activities of an object recorded by the sensor devices S (Chen *et al.*, 2005). Thus, S could be normalizing to x and y position values using their respective mean values as (μ_x, μ_y) and corresponding standard deviations as (σ_x, σ_y), which defined as:

$$S(x_i, y_j) = \left[\left(t_1, \left(\frac{S_{1.x}-\mu_x}{\sigma_x}, \frac{S_{1.y}-\mu_y}{\sigma_y}\right)\right), \ldots, \left(t_n\left(\frac{S_{n.x}-\mu_x}{\sigma_x}, \frac{S_{n.y}-\mu_y}{\sigma_y}\right)\right)\right] \tag{1}$$

The advantage of SDM similarity function is a dynamic and scalable computational technique for large-scale sensor data. The approach helps to develop behavioral patterns-based analysis supported by a corresponding time and sensor locations. The sensor data records at each sensor device are a pair of two parameters, which is pertinent to detect where such abnormality or faults happened. It is the other way of evaluating the intelligent system performance to improve the smart space applications. The analysis of such dimensional data patterns needed for system reusability and in general fostering performance optimizations. It is also gaining in the target area for ubiquitous and pervasive computing, which focussed on a proprietary intelligent system into "real living spaces," such as the residential home.

3.2 Sensor data characteristics

The sensor data characterization is essential for classifications and other further analysis, which gives a meaningful understanding and standards to category's events as their common properties and data trends. It requires efficient and real-time processing techniques to avoid a problem that arises from its massive volumes of possibly uncertain nature and complexity. The analysis of the active (or live) sensor data is needed to execute in one pass of the data, which mean that the data type is typically not often possible to store into the entire data set.

SE is a smart place of the real world, which equipped with sensors, actuators and computing components that generate massive sensor data. The data types (volume, velocity, variety, etc.) of sensor records characterize as its independent sources and object behavior. It is essential to extract valuable information to optimize the intelligent system, including computer application's performance and safety that affect people or users' daily activities (Szewcyzk *et al.*, 2009). For example, the activity of walking or cooking in the smart place, recognize as actuator's behavior in terms of objects' moments sequential events of the sensor data. The sensor signals recorded events sequential properties against time, sensor locations and object's occupation in a smart space (Diane and Lawrence, 2011). The sequence of k event types is called a (k)-sequence. It refers to the subsequence consisting of the ith to the jth event types of k-sequence $\overrightarrow{S}$ as $\overrightarrow{s}(i,j)$, with $i \in [1, k-1]$, $j \in [i+1, k]$, and the ith event type in the sequence $\overrightarrow{S}$ as $\overrightarrow{S}[i]$ that assume. The assumption of any event type $\overrightarrow{S}[i]$ in $\overrightarrow{S}$ is needed to follow its sequence of the $\overrightarrow{S}[1 : i-1]$ (Tao, 2012; Juan *et al.*, 2010).

3.3 Sensor data models for pattern mining

A fundamental aspect of data modeling and pattern development is a technique to prepare sensor data for knowledge extraction and fault (abnormal behavior) diagnosis of the intelligent systems. It is a sequential arrangement of sensor data according to object's behavior or properties signal that capture or record by the sensor devices in the system. The processes of event sequence conceptualize on sensor locations, and trajectory records time. The path for each event's movement shapes the behavior of the object's data that need to analyze accordingly (Dhaval *et al.*, 2009) using different data analytical tools. In the past, various techniques implemented to visualize data patterns, which include Bayes Gaussian (a probabilistic approach) (Jaakko *et al.*, 2008; Rainer *et al.*, 2008), neural networks (Ahmad *et al.*, 2011; Fernando *et al.*, 2005). However, the sensor data exploration process is dynamic and steps wisely, which need the advanced analysis to understand and clearly define signal behavioral patterns. SDM techniques are capable, adaptable and make the analysis to more understandable, generic and informatics as the intelligent system contexts (Fan *et al.*, 2009). The data processing conceptual model, as it is shown on Figure 1, is a basic flow chart to recognize activities or events as the contexts of supervised learning techniques approaches that relying on the unsupervised sensor data. As Tˆam and Bent (2006) proofed, the process depends on labeled data that the activity recognition would limit to the activity learning the recognition of sensor data.

Patterning mining could not be the end of data processing, which extend until knowing or defining the data behavior as normal and abnormal patterns. The ABP mining is more about the process of defining the pattern that deviates from the normal context of the given sensor data. Then after, based on the detected pattern, we extract valuable information that can changed to the knowledge for the improvements of intelligent system performances. Sensor data are always complex and context sensitive, which do not have a common or predefine definition and standards (Paul and Richard, 2001). Such challenges are essential issues in the implementation of SDM to develop a sequential pattern as pervasively of the sensor data analysis gain to the new patterns. Furthermore, properly modeling data are important to describe the behavioral correlation of event records against time and locations to explore the existing situations of the systems.

3.4 *Sensor data model descriptions*

The model description is the representation of data behaviors and characteristics in various ways by having time and locations inferences, which is important to detect and isolate data behaviors as the domain contexts. The models can be developed either as the underlying of the physical system as a set of devices establishing a relational coordination in which characterized as magnitudes of trajectories to apply deterministic behavioral patterns. In some case, the process might be a complex and model description also challenging. It needs to support a priori probabilistic models, which support to apply linear regression (Diane and Lawrence, 2011) and Gaussian models (Szewczyk *et al.*, 2009) techniques. The sample data pattern similarities such as the trajectory length to their respective locations and behavior are developing as Figure 2. The patterns (P_i) characterized by three distinct colors as P_1, P_2, and P_3, which referenced by the sensor locations S_1, S_2 … S_i and (L_i) are the trajectories' lengths.

As the sensor data behavior change, the pattern will also change, which is interesting to the approach of SDM techniques that needs to activity recognitions as the domain contexts (Chen *et al.*, 2005). The techniques of Naïve Bayes classifiers as a DM approach vitally essential to identify the activity, while the events correspond to the greater probability set of sensor values. Figure 2 showed if, someone (an object) wants

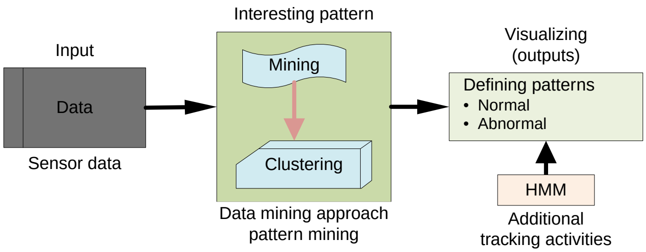


Figure 1.
Proposed pattern mining components model

to go from "S_1" to "S_6", the moment will record in different sensors as its paths, which causes a certain deviation sensor data. Objects' trajectory data are discriminating in terms of time, magnitudes, with direction and sensor device locations, which are fundamental factors for pattern behavior detections to define as nor normal and abnormal. Because the sensor captured data are integrated or described in such parameters (Table I), which support to verify the activity recognition model to its proper location positions.

Sensor data are not always labeling. In some case, it may have uncertainty. As it is showing in Table I, the records have some unlabeled data on the second row of the third column. In such case, the unlabeled records can be defined based on the original labeled data by performing activity recognitions of model descriptions using the bootstrapping learning technique (Peng *et al.*, 2007). The approach is using basic radiative transfer equations and takes advantage of unique multichannel distributions of sensor emissivity to derive sensor behavioral pattern of the unlabeled data (Ariel *et al.*, 2012). An inherent assumption of a bootstrap algorithm is operating in its optimal performances for the large-scale sensor data. The sensor data in a sequence of events "e" are described as $e = \{t, id, l\}$. Where "t" is a timestamp, "id" is the sensor "**ID**", and "l" is the activity of the label. The **ID** associated with the sensors in the system or the space that refers to a location tag of "L", which used to facilitate to transform the activities' patterns in the specified nodes. The activities are donated by "$\boldsymbol{a}$". It represents as $a = \{E, l, t, d, L\}$, where E are sequential patterns of n sensor device of events "e" patterns. It is $\{e_1, e_2, \ldots, e_n\}$. t and d are the start time and duration of the activities, respectively as shown on Figure 2 of the sensor data features model, and the graphic components are the design of sensor data patterns.

The pattern S_i of the ith sensor device is designed to the specifying locations to record the moving object sequence actions. The sequence describes as $S = \{s_1, s_2, \ldots, s_n\}$ for all "n" sensor device nodes, and the records are coordinating as its position (location) of "S_i" and time "t_i". Therefore, the locations s_i as Union (U) are defined as

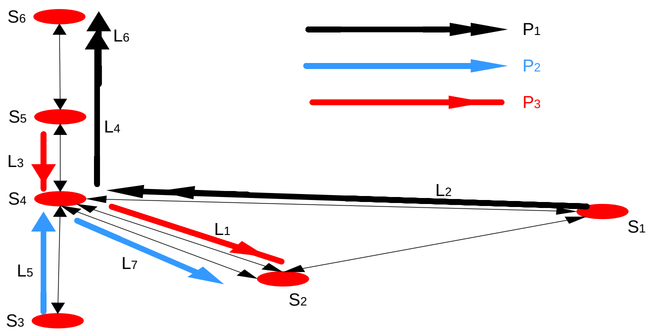


Figure 2. Sensor node based modeling representation of the sensor data set observation patterns

Table I. Example sensor data records as walking (W) of W004, W030, and W015 denote sensor IDs

Timestamp (t) sensor	ID (id)	Label (l)
17/10/2011 06:00	W004	Conference room
17/10/2011 06:01	W030	None
17/10/2011 06:02	W015	Reception

$U = \{u_1, u_2, \ldots, u_n\}$ from its sub-components, which are the set of activities in the smart space where those sensors can detect object behaviors along their paths. To isolate the detected abnormal behavior, we process as targeted activities (T) defined as $T = \{t_1, t_2, \ldots t_n\}$, which represent events sequence of instants starting at the time t_0. The records could involve about the object trajectories, including come into, walk around and leave the area during the given time t. When an object positioning on a sensor s_i decides to move to a new position that the data recorded by the respective sensor s_j at its new location and the destination. It implies $s_i \rightarrow s_j$ that denoted as the direction of transition for the objects. It measures by a transition relative distance matrix, denoted as Ω that hidden behind these sensor vectors "S" (Fan *et al.*, 2011). The transition distance between those reachable nodes should be introduced into the trajectory searching algorithm, which might not be necessary to reflect the real distance that can be a simple transition relative distance (Jan, 2001):

Definition 1. transition relative distance $d_{si \leftrightarrow sj}$ denotes that the relative distance of moving from s_i to s_j, both of which should be physically reachable without triggering any other nodes. The model $d_{si \leftrightarrow sj}$–0 represents that there is no chance, and the smaller value indicates, the nearer distance between them. We set 0 to those physical unreachable relationships between sensors and calculate transition relative distance using a standard distance value, which divided by the real distance measured from the deployment of the sensors.

3.5 Sequential data mining-based sensor data synthesizing

Sensor data synthesizing is the process of annotates the data to define the key components for a clear understanding and look for relationships between the events, which collected from various sources. It is a way to know data characteristics, interactions and interrelations to explain the sensor data in which contextual conditions, actions, and consequences affect the data behavioral patterns. Therefore, data synthesizing considers the different perceptions and standpoints of sensor data and the multiple and diverse patterns of connections for abnormal behavioral pattern detections. The process is pertinent for many reasons, which includes: first optimizations and makes efficient in the intelligent system via SE. Second, exploring the diversified application of sensor data, which is creating new user models that lead to easy access and interaction the objects as needed. Third, it is a systematic approach to searching the hidden knowledge from the sensor data for a better of the decision-making process. Fourth, to personalize the smart systems for which to analyze the user's interaction with a system and developing cognitive models that aid in the design of user interfaces and interaction mechanisms. It is human-centric that provides context information on location, trajectories and the behavior of the intelligent systems (Debraj *et al.*, 2012). Fifth, it is essential to demystifying ABP detections and isolations as the domain contexts. Sensor data Annotation is also a fundamental task to standardize the sequences and indexing process, including the sensor ID, time and activities' initial and destination as it is shown in Table II.

Table II.
Sensor data on moving objects

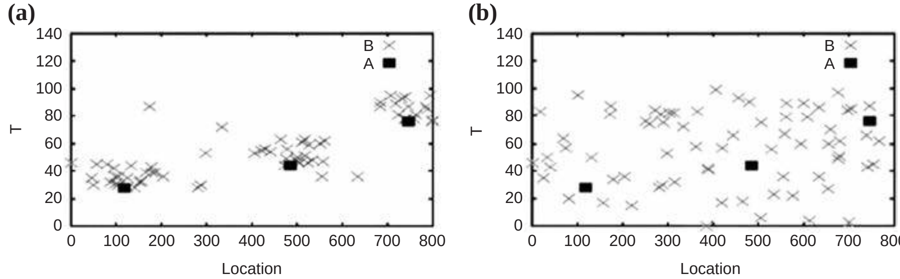
Date	Time	Sensor ID	Object ID
2012/04/30	10:56	S1	O1
2012/04/30	10:57	S2	O2

3.5.1 Sequential pattern mining. Sensor records are sequential, and continuous flow of signal's data streams in the sensor devices. The synthesize and annotate of sensor data refine the data behaviors and trajectories, which support behavioral patterns detections using SDM algorithms (Thomas *et al.*, 2006). The process begins with the conceptual understanding of events (*E*), and the pattern defining as $E = \{e_1, e_2, \ldots, e_n\}$ to the sensor data sets "*D*". Each event of the sensor data consists of the field's event-ID (id), and the instance types (a_i, b_j) as it showed in Figure 3. Therefore, the Pattern (*P*) is further categorized as normal (P_n) and abnormal (P_a) in reference to location and time of objects trajectories.

As it is shown on Figure 3, the two events of "*a*" and "*b*" correlations are symbolized by "×" and "square black box", which represent the important interactions of the data behaviors. The two correlations points as data behavioral patterns tell about events factors dependency via time and locations. Then after, the exploration techniques are exhibited based on the data patterns, which might deviate from the normal positions or not. The way to synthesize these events: first, it classified the unstructured data of its similarities that formulating as the similarity patterns (sim_P) and a minimal threshold ($sim_{\min} e_i$ *and* e_j) of the events' pattern. Then the events' relation score is defined as ($sim_P (e_i, e_j)\ sim_{min}$), which helps to measure the distance variations between the two events. An event sequence $E = \{e_1, e_2, \ldots e_n\}$ is an ordered event, where an event E_i ($1 \leqslant i \leqslant l$) is a moment (or transformation) from its start to end. "*l*" is the number of events in a sequence of a trajectory length. Each length might (not) have a specific attribute, such as timestamp, denoted as $e_i \times time$ that registers the time when the trajectories are executing. As a generalized convention, sequence $E = \{e_1, e_2, \ldots, e_n\}$ and $e_i \times time < e_j.time$ are represented as $1 \leqslant i \leqslant j \leqslant l$.

A sequence E with length l is an l-sequence, denoted as $len(E) = l$ and the ith sequence denoted as $E[i]$. An event can occur at most once in event sets, and can occur multiple times in various event sets in a sequence, such as sequence E, β, etc. As the behavioral pattern variations, one can be super sequence than another for a given sensor data. Sequence sensor data set is a set of two-tuples (sensor ID and events (*E*)). A tuple (S_ID, *E*) in a sensor data is said to contain a sequence γ that γ is a subsequence of E. For example, Table III show's event sequence of sample sensor data sets at the minimum threshold ($\theta = 1$) signal detections as the count value of $a > \theta$, $b = \theta$ and $c < \theta$ for each sensor device.

The sensor detects the object as the sequence events, where each signal in the reading time and location correspond to sense objects as its positions and functionality. If the object is a moving object, the trajectory synthesizes as events order against the increasing trajectories-time that corresponds to a sequence sensor data. The trajectory length measured at time series for a sequence E of the intelligent systems. The sequence describes



Notes: (a) Follows "a"; and (b) "b" is independent of "a"

Figure 3. The two events of "*a*" and "*b*" sequences

as the fraction of total sensor records that support the event sequence. The sensor data behavioral pattern sequence analysis of a moving object or detection trajectory is supported a minimum of two sensor devices, which helps to detect events that are greater than or equal to the minimum threshold ($e_i \geqslant \theta$). Based on this concept, the events' sequence pattern is sorted from largest to least detections as Table IV.

Suppose the event (a) detection is supported by sensors 1-4, (b) supported by 1, 4 and 5, (a, c) by 2 and 3, and (a, b) by 1 and 4. Therefore, the sensor data behavioral pattern is varying as the detected event sequences and trajectory lengths as shown in Table IV. The advantages of events sequences are also important to define the context of the data, which are capable to build the proper pattern in the sensor data along with their variations. From the intelligent system context model, the process of different sensor data may have diverse patterns due to the variation of space, time or activities. Such multiple patterns can be addressed by providing multiple thresholds of the data behavior and sequences. The SDM approach of Sequential Pattern Mining (SPM) and the physical environment contexts are important to the system able to detect the ABP value (Yan *et al.*, 2008). That is sensor "*i*" is (S_i) as a group of "*n*" point's instances or activities of "*a*", and it is defined as:

$$s_i = \{a_i, a_2, \ldots, a_n\} \tag{2}$$

Based on these definitions, event sequences of pattern (a, b) distance variation can be also computed in regardless of the instances' order variations or incompleteness of sensor records. However, the derived pattern (a) event sequence could comprise *n* variations for the activities, which denoted as a_i. Furthermore, if there are discontinuous events or records, the pattern would be a discontinuity of zero. Thus, the sequential similarity function is vitally significant to define the pattern behavior, which is imperative to have common features between instances in terms of time series aspects. The distance can be measured using different techniques, including DM, HMM and

Table III. Sample events pattern sequence

S_ID	Sample events	Transformed events sequences	Sorted events sequences: Sensor ID	Sorted events
1	a, b	< (a, b) >	1	< (a, b) > < (c) >
1	c	(c)	2	< (a, c) >
2	a, c	< (a, c) >	3	< (a, c) >
3	a, c	< (a, c) >	4	< (a, b) >
4	a, b	< (a, b) >	5	< (b) > < (c) >
5	b	(b)		
5	c	(c)		

Table IV. Large detections sets of event sequence

Large detection sets	Mapped to
(a)	1
(b)	2
(a, c)	3
(a, b)	4
(c)	5

Dynamic Time Warping (DTW) (Chen *et al.*, 2005). DTW is needed to measure the dimensions of time variations between instances. The mathematical formulation of the two instances distance computed based on Hausdorff distance measurement of following equation as shown in (Table V):

$$D(s_i, s_j) = \max\{d(s_i, s_j), d(s_j, s_i)\} \tag{3}$$

where $d\ (S_i,\ S_j) = \max_{a \in Si},\ \max_{a \in Si}\ \|a - b\|$

The measurements of the path or distances between "*a* and *b*" are normalized by 26.2368 of the dimensional transformations from its centers (0, 0), which can be computed as:

$$\tilde{m}(s) = \frac{1}{n}\sum a_i \tag{4}$$

The sensor data similarity along with contextual information is an abstraction step to segment and labels the real-valued time series into similar subsequences. The normalized computed distance measurement revealed the relational pattern that defines the sensor data features, which is important to discover the sensor data behavioral patterns clustering.

3.5.2 Behavioral pattern clustering. Behavioral pattern clustering is a technique of categorizing sensor data as of their similarities and interactions to achieve the goal of keeping track of intelligent systems performances. The advantages of clustering based SPM is to predict the object behavioral positions in terms of locations and time to detect abnormal behaviors or faults (Bhatia and Deepika, 2013; Elad, 2004). The algorithm, such as a *k*-means use for the set of events sequences or trajectories as clusters (*C*), are $\omega = \{C_1,\ C_2, \ldots, C_k\}$. Whereas, the trajectories (T_r) clusters are a sequence of multi-dimensional points, which is donated as $T_{ri} = (a_1,\ a_3, \ldots, a_j)$ a subject to be $a_i\ (1 \leqslant \mathrm{j} \leqslant len_i)$ is a dimensional point. Where "len" is the length of the trajectories or number of events "*N*" (activities). The length "len" of a trajectory differentiates as $T_{rc1},\ T_{rC2}, \ldots, T_{rCk}$ $(1 \leqslant \mathrm{C}_1 < \mathrm{C}_2 <, \ldots, < \mathrm{C}_k \leqslant \mathrm{len}_i)$ of the sub trajectory as the Definitions of 2-4:

Definition 2. Sensor data pattern is the events or signals sequential design that models to understand its behaviors as nor of normal and abnormal behavior patterns. The data set $A = \left\{a_i \in \mathfrak{R}^d\right\}_{i=1}^{n}$ represented in each activity in a pattern of *d-D* (dimensional) set of data that associates

	*S*1	*S*2	*S*3	*S*4	*S*5	*S*67	*S*7	*S*8	*S*9
*S*1	0	23.6643	18.8812	19.8620	20.9105	20.0873	20.5244	22.0284	22.0681
*S*2	23.6643	0	20.0000	19.8746	20.1928	21.7256	21.2544	21.7658	22.5389
*S*3	18.8812	20.0000	0	19.6342	17.0514	19.2224	17.6564	19.9060	19.5320
*S*4	19.8620	19.8746	19.6342	0	19.9437	18.6145	19.8305	20.9344	21.5058
*S*5	20.9105	20.1928	17.0514	19.9437	0	21.0891	19.3520	20.1370	20.3408
*S*6	20.0873	21.7256	19.2224	18.6145	21.0891	0	22.1190	23.0922	22.0567
*S*7	20.5244	21.2544	17.6564	19.8305	19.3520	22.1190	0	23.4947	22.4332
*S*8	22.0284	21.7658	19.9060	20.9344	20.1370	23.0922	23.4947	0	24.9048
*S*9	22.0681	22.5389	19.5320	21.5058	20.3408	22.0567	22.4332	24.9048	0

Table V. Hausdorff distance measurement based sample data correlations with their class label $\omega(a_i) \in \{\omega_a(ABP), \omega_n(normal)\}$. Therefore, our assumption is $\forall a_i, a_j, \omega(a_i) = \omega_{P_a} \Lambda \omega(a_j) = \omega_{P_n} \Rightarrow f(a_i) \geqslant f(a_j)$, i.e., any P_a factor must be greater than all P_n factors. The predicted ABP set as (P_a), which represented as $\{a{:}\omega(a) = \omega_{Pa}\}$.

Definition 3. *ABP is* defining and computationally feasible for a large set of sensor data, which is a process of fault or unusual behavior detections based on k-mean algorithm and distance measurements. The techniques involve the hidden interactions of the events and data similarity of pattern frequency. Thus, the procedural prediction of an event e_1 is followed by its nearest neighbor e_2.

Definition 4. The event e' follows N of a different event e, denoted by $e \rightarrow Ne'$, subject to e' is in the sensor data Ne' of e patterns. A simple sensor event, neighborhood of an event e to capture the follow predicate (p), distance (R) and time interval (T) can be defined as:

$$N(e) = \{p \mid p \in D \wedge distance(e.location, p.location) \leqslant R \wedge (p.time - e, time) \in [0, T]\}$$

The significance measure of $f_1 \rightarrow f_2$ the sequential pattern describes as the number of events of the type f_2 that follow events of type f_1. It is essential to detect sequence in a large-scale sensor data within any independent distribution system (Yan *et al.*, 2008). It is a systematic approach to identifying the proper sequential even patterns than count them. The clustering technique for the sequential pattern is a systematic data model, which provides a clear understanding of sensor data similarity of neighboring node pattern and instances categories. One or more instance could be in one pattern, but not the reverse. For instance, one object trajectory can have different instances, such as slow, normal, fast walking. However, the pattern is the same that is the object walking. Therefore, the process of clustering is essential to define the abnormal behavior patterns from the object trajectory patterns, which support the instances of the domain contexts (Piciarelli and Foresti, 2006). The proposed pattern clustering algorithm or technique is defined as follows:

Algorithm: Pattern clustering method

Procedure Cluster (P)
  ►At the beginning, each pattern is considered as a cluster
    C = p
    Compute proximity matrix, m
    repeat
    Sim = max(m[a, b]) $\forall_{a,\ b} \in C$
    if sim > θ **then**
      Merge a, b
    end if
    update m
  until sim > θ
  return C
end procedure

Furthermore, the trajectory (T_r) partitioned can be classified at every event points (P_t). The partition represents a line segment between two consecutive characteristic points as the set of (P_{pari}–l) line segments point clusters (P_{tci}) and partition (P_{pari}) as {p_{tc1}, p_{tc2}, …, p_{parc1}, p_{parc2}, …, } of the sensor data.

3.5.3 Distance measurement based behavioral pattern mining. Distance-based clustering is a pairwise similarity and distance variation of the events, which perform as the point distance measurements between two points. The process of distance computing consists of a sequence of activities $a_1, \ldots, a_n$ at time $t(l, n)$ that can be explored using different distance measurements. The distance between the pairs of activities or locations measures as its corresponding points of $\boldsymbol{a}$ and $\boldsymbol{b}$ using the Euclidean distance measurements. The pairs of the sensor record are the proportion time length, i.e. $a_{1,\ldots,n}$ and $b_{1,\ldots,n}$ of $E_{uclud}(a_{1,\ldots,n}, b_{1,\ldots,n})$, which defined as the sum of activity's distances in a given time t interval. It described as:

$$D_{Euclid}\left(a_{1,\,\ldots,\,n}, b_{1,\,\ldots,\,m}\right) = \sum_{i=1}^{n} \left\| a_t, b_t \right\| \tag{5}$$

The Euclidean distance between the two activity points defined as:

$$\left\| a_t, b_t \right\| = \sqrt{(a_{t.x} - b_{t.x})^2 + \left(a_{t.y} - b_{t.y}\right)^2} \tag{6}$$

Furthermore, the noisiness and distortions of the sensor data may affect the alignment of the pattern, which the Euclidean distance measures could not capture the inherent distance of the trajectories. Therefore, applying various alignment distance measurements such as DTW is paramount. To compute the distance between two points is the systematic approach to permitting the comparison of the Euclidean distance of activities with different time lengths (Karthika and Sumathi, 2012). The method is employed to align the two trajectories as of the overall distance minimization process and the dynamic computation of DTW of (n, m) time defined as:

$$DTW\left(a_{1,\,\ldots,\,n}, b_{1,\,\ldots,\,n}\right) = \left\| a_n, b_m \right\| + \min \begin{cases} DTW\left(a_{1,\,\ldots,\,n-1}, b_{1,\,\ldots,\,m-1}\right) \\ DTW\left(a_{1,\,\ldots,\,n-1}, b_{1,\,\ldots,\,m}\right) \\ DTW\left(a_{1,\,\ldots,\,n}, b_{1,\,\ldots,\,m-1}\right) \end{cases} \tag{7}$$

Where $a_{1,\ldots,n-1}$ and $\mathrm{b}_{1,\ldots,\mathrm{m}-1}$ are the sub trajectories of $a_{1,\ldots,n}$ and $\mathrm{b}_{1,\ldots,\mathrm{m}}$ respectively that covers the time points 1 *to* $n-1$ sequences. For Metrix sensor data sets, Edit Distance on Real sequence (EDR) measurement is more powerful to compute the sequence pairs of numerical values of points, which is significant to define proper matching between instances:

Definition 5. The pair of activity vectors a_{i} and b_j is from two trajectories of a and b. The corresponding matching represents *match* (a_i, b_j) = *true iff* $|a_{ix}, b_{jx}| \leqslant \theta\ b_{jx}$ *and* $|a_{iy} - b_{jy}| \leqslant \theta$, where θ is the matching threshold. The two trajectories a and b of length with their corresponding length n and m *EDR* (a, b) is the number of inserts or replace or delete operations to shift a into b. Therefore, EDR distance of the two trajectories "a^n" and "b^m" is defined as:

$$EDR(a^n, b^m) = \begin{cases} \infty & if \quad m = 0 \; and \; n = 0; \\ n.c_{ins,} & if \quad m = 0 \; and \; n \neq 0; \\ m.c_{del,} & if \quad m \neq 0 \; and \; n = 0; \\ & \min \{EDR(a^{n-1}, b^m) + c_{del,} \\ & EDR\left(a^{n-1}, b^{m-1}\right) + c(a_n, b_m), \\ & EDR\left(a^n, b^{m-1}\right) + c_{ins}, \quad \text{else}, \end{cases} \tag{8}$$

Where "C" is the sub cost function of the two activities (of a and b). Therefore, $C(a_n, b_m) = C_{match}$ if $dist\ (a_n, b_m) \leqslant \theta$; $c(a_n, b_m) = c_{mismatch}$ $if\ dist\ (a_n, b_m) > \theta$ and. c_{match}, $c_{mismatch}$, c_{ins}, c_{del} are the cost of the match, mismatch, insertion, and deletion respectively, subject to $c_{ins} = c_{del}.\theta$ symmetry to the matching threshold that quantizes the uncommon to distance C_{match} or $C_{mismatch}$. If the sensor data is the triangle inequality types, the Longest Common Sub Sequence measure (LCSS) (Jae-Gil *et al.*, 2007) is more appropriate. LCSS distance measurement needs parameterization of the two points as δ and ε to match with their respect time and space, and the computation is defined as:

$$LCSS(a_{1,\ \ldots,\ n}, b_{1,\ \ldots,\ n}) = \begin{cases} 0 & \text{if} \quad n = 0 \vee m = 0 \\ 1 + LCSS(a_{1,\ \ldots,\ n-1}, b_{1,\ \ldots,\ m-1}) & \text{if} \quad |n-m| \leqslant \delta \\ \max\{a_{1,\ \ldots,\ n-1}, b_{1,\ \ldots,\ m},\} & \wedge \|a_n - b_m\| \leqslant \varepsilon \\ LCSS(a_{1,\ \ldots,\ n}, b_{1,\ \ldots,\ m-1}) & \text{otherwise} \end{cases} \tag{9}$$

For local distance or density or density based measurements, Minimum Boundary Rectangles (MBR) approach is more effective and pertinent (Stephen and Ernest, 1997). It is a method to approximate line segmentations between trajectory distances. The two boundaries of B_1 and B_2 of the MBR of the line segments are L_1 and L_2, respectively. The local distance based on the MBR of $D_{\min}(B_1, B_2)$ computed as:

$$D_{\min}(B_1, B_2) = \sqrt{(\Delta(B_1.[x_1, x_u], B_2[x_1, x_u]))^2 + (\Delta(B_1.[y_1, y_u], B_2.[y_1, y_u]))^2} \tag{10}$$

The distance between two points defined as:

$$\Delta([l_1, u_1], [l_2, u_2]) = \begin{cases} 0 & & [l_1, u_1] \cap [l_2, u_2] \neq 0 \\ l_2 - u_1 & if & u_1 < l_2 \\ l_1 - u_2 & if & u_2 < l_1 \end{cases} \tag{11}$$

Hausdorff distance measures based on a weighted sum defined as:

$$D_{Huasdorff} = w_{\perp}.d_{\perp} + w_{\parallel}.d_{\parallel} + w_{\theta}.d_{\theta}$$

Where $w_\perp$, $w_\parallel$ and w_θ are the weight of the components of the aggregate perpendicular distance ($d_\perp$) of the two separated trajectories ($d_{\perp a}$ and $d_{\perp b}$). The aggregate parallel distance ($d_\parallel$) of two captured between trajectories, and the angular distance (d_θ) of the two trajectories orientation differences as it is shown on Figure 4(a):

Definition 6. $d_\perp$, $d_\parallel$ and d_θ is the distance functions in clusters line segments of L_1 and L_2, which defined as the Equation (12). The Euclidian distance between $d_{\perp a}$, $d_{\perp b}$ are two perpendiculars distances between L_1 and L_2. $d_{a\parallel}$, $d_{b\parallel}$ are the two parallel distances between the two lines, and θ is the angle between L_1 and L_2.

The distance function between two trajectories line segments of L_1 and L_2 are defined as:

$$\begin{aligned} d_\perp &= \frac{d^2_{\perp,a} + d^2_{\perp,b}}{d_{\perp,a} + d_{\perp,b}} \\ d_\parallel &= \min\{d_{\parallel,a}, d_{\parallel,b}\} \\ d_\theta &= \|L_2\| \times \sin\theta \end{aligned} \tag{12}$$

The sub-cluster of the sensor data (trajectory) partitioning is also a line segment of two points (Figure 4(b)) that belong to the same cluster closed to each other according to their distance variations. The trajectory line segments of $l_1^{'}, l_2^{'}$ are the approaches to support the efficient trajectory clustering, and it is defined as:

$$D_{LL}\left(l_1^{'}, l_2^{'}\right) = \min_{a_i \in l_1^{'}, b_j \in l_2^{'}} \|a_i - b_j\| \tag{13}$$

3.6 Comparative analysis

Sensor data are complex and continual time series or streaming data, which is demanding powerful analytic tool to develop its behavioral pattern toward faults or abnormal behavior detection. In the past various approaches were proposed for analysis, such streaming data, via HMM and Bayes classifier. The HMM is implementing to measure the distance variations by modeling the data as their similarities and hidden interactions using Hilbert Scanning Distance – HSD, Kullback-Leibler distance (KLD), heuristic distance, and others. The techniques are more effective and successful in speech recognition, time series prediction, and document and image classification than sensor data analysis (Jianping *et al.*, 2010). Whereas, the Bayes distance measurement is also a probabilistic approach. Nutshell, the sequential data analysis using these knows techniques are complex and subject specific.

In either of HMM, various techniques and Bayes based Hidden Markov process consists of a state transition process. These are the hidden data behavior and the observable process that determined by the underlying the state transition. HMM (λ) is

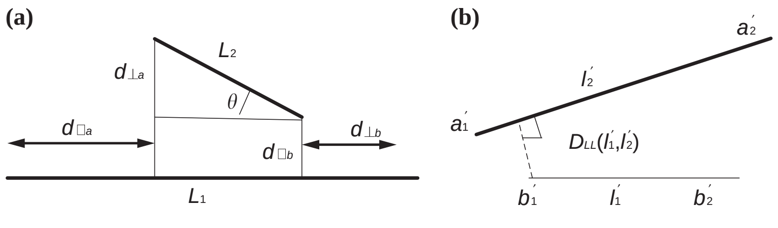


Figure 4. The three components of distance function for line segments

the descriptions of data initial distribution (π), transition matrix (A) and the output distributions (B), which described as $\lambda = (\pi, A, B)$. HMM-based distance measurement is the process of data transit from its initial to the prospective output as a model. The output distribution or B described by the action of emitting as $b_i(x)$, which provides an emission matrix indicating the probability of events of x from a hidden state i as a normal distribution. Therefore, as per the hidden similarities, the discrete HMM model charactering as the set of N hidden states in the model of $S = \{s_1, s_2, \ldots, s_N\}$ sequences. The state transaction probability matrix is:

$$A = \{a_{ij}\} \quad 1 \leqslant i,j \leqslant N \tag{14}$$

where: $a_{ij} = p\{q_{t+1} = s_j | q_t = s_t\}$

$$0 \leqslant a_{ij} \leqslant 1. \quad \sum_{j=1}^{N} a_{ij} = 1$$

The set of M observable events per state is, therefore, $E = \{e_1, e_2, \ldots, e_M\}$ and the observation events, probability matrix is $B = \{b_i(x)\}$ where $1 \leqslant j \leqslant N$, $1 \leqslant x \leqslant M$, thus:

$$b_j(\mathrm{x}) = p\{e_i \; at \; t | q_t = s_j\}, \text{ subject to } 0 \leqslant b_j(x) \leqslant 1, \sum_{i=1}^{N} b_j(x) = 1 \tag{15}$$

Therefore, the initial state probability distribution is:

$$\pi = \{\pi_t\}, \quad 1 \leqslant i \leqslant N \tag{16}$$

where:

$$\pi_t = p\{q_1 = s_t\} \tag{17}$$

$$0 \leqslant \pi_t \leqslant 1, \quad \sum_{t=1}^{N} \pi_t = 1$$

IV. Experiment and discussion

The rapid development of sensor technology has paved the way for SE to provide intelligence sensor-based services, which contains several and advanced interconnect smart devices. The intelligent environment is, therefore, has the abilities of perception, cognition, sensations, reasoning and anticipation about a user's activity, which provide proper reactions (Zhonghong, 2008). In parallel, sensor data generated by such a sophisticated and intelligent system are increasing ever than before, which demand advanced and scalable analytic tool. DM is an effective tool to manipulate in understandable ways as to develop sequential data patterns. Mining sequential patterns as SDM technique has become an important task with broad applications, such as abnormal behavioral pattern detections or fault diagnosis. The objective of detecting and isolating abnormal behavioral pattern is to understand the factors that affect SE performances, which offers an intelligent sensor-based service. The technique implemented DM-based clustering and distance measurements on the sample sensor data.

The performances of the proposed approach evaluations showed on the subsequent analysis, which produced results that provide more clear and novel explanations about the need for SDM-based critical analysis for sensor data. The effectiveness and efficiency of the approach are tested in iterative and extensive experiments using Center for Advanced Studies in Adaptive Systems (CASAS) sensor data (CASAS, 2010). Moreover, SDM-based behavioral pattern mining discussed in terms of its tremendous applications, including fault diagnosis and performance optimizations for the better functionality of smart spacing and technology. CASAS data obtained from the CASAS smart home database by subscribing and downloading from their web site. We used 43.8 MB sample data that having 22 attributes (sensors index) with 729,000 records. To explore this data as an SDM approach, we used MATLAB and RapidMiner application tools. The approach is a step-by-step sensor data projection and pseudo-projection techniques. We use nine attributes (as random selections) and their corresponding 307 records as a training sensor data as shown in Figure 5 and then scale out for large-scale sensor data analysis.

In an SE, object moments are recognizing and synthesizing at the corresponding time of the records as a cluster of trajectories, which is worth noting that the design of the sensor devices used for various applications. The sensor data tabulated in the form of $n \times p$ the data matrix $X = (e_{ij})$. Where "n" is for the number of records (raw data) and "p" for the parameters (S_1-to-S_9 of the sensor devices) as showed in Figure 5. The set of objects clustered in terms of attributes or events that we seek as a collection of k-mutually exclusive and exhaustive subsets of X, say, $C_1, \ldots, C_k$. The clusters are an agglomeration of normal and abnormal behavioral patterns of the events. The clusters' elements and the sum-of-squared-error criterion are defined based on the events and patterns distance measurements against their nearest neighbors. The allocation of events to the respective groups is carried out according to minimum distance and the cluster centroid calculated values. The minimum distance computations of cluster-centroid are performing until no change is possible. However, the numbers of clusters (k) are not static that vary as the user desires. In this study, we computed various iterative clustering tests to minimize the risk of data masking and data swamp for proper ABP detections. Among the various outcomes, for "k" values of 2 and 5 are clear and understandable graphic representations showed on Figure 6(a) and (b).

Note that the process of clustering is a representation of patterns in respect of each event being closest to its group cluster-centroid. We choose hundred random starts ("replicates", 100) by picking events at random to serve as the initial events ("start", "sample"). The control phrase ("maxiter", 500) increases the allowable number of iterations. Each chosen clusters obtained from the hundred replications with **idx** (index of x (activities)), which indicating cluster membership for the n events. It evolves the cluster centroid "c", the sum deviations of *sumd* gives the within-cluster sum of

ExampleSet (307 examples, 0 special attributes, 9 regular attributes)

Role	Table Index	Name	Construction	Type	Statistics	Range	Missings
regular	0	S1	S1	real	avg = 0.419 +/-	[0.000 ; 4.000]	0
regular	1	S2	S2	real	avg = 0.471 +/-	[0.000 ; 4.000]	0
regular	2	S3	S3	real	avg = 0.539 +/-	[0.000 ; 3.500]	0
regular	3	S4	S4	real	avg = 0.533 +/-	[0.000 ; 3.500]	0
regular	4	S5	S5	real	avg = 0.306 +/-	[0.000 ; 4.000]	0
regular	5	S6	S6	real	avg = 0.406 +/-	[0.000 ; 3.500]	0
regular	6	S7	S7	real	avg = 0.440 +/-	[0.000 ; 3.000]	0
regular	7	S8	S8	real	avg = 0.453 +/-	[0.000 ; 4.000]	0
regular	8	S9	S9	real	avg = 0.412 +/-	[0.000 ; 4.000]	0

Figure 5.
Sensor data example sets with detail attributes descriptions

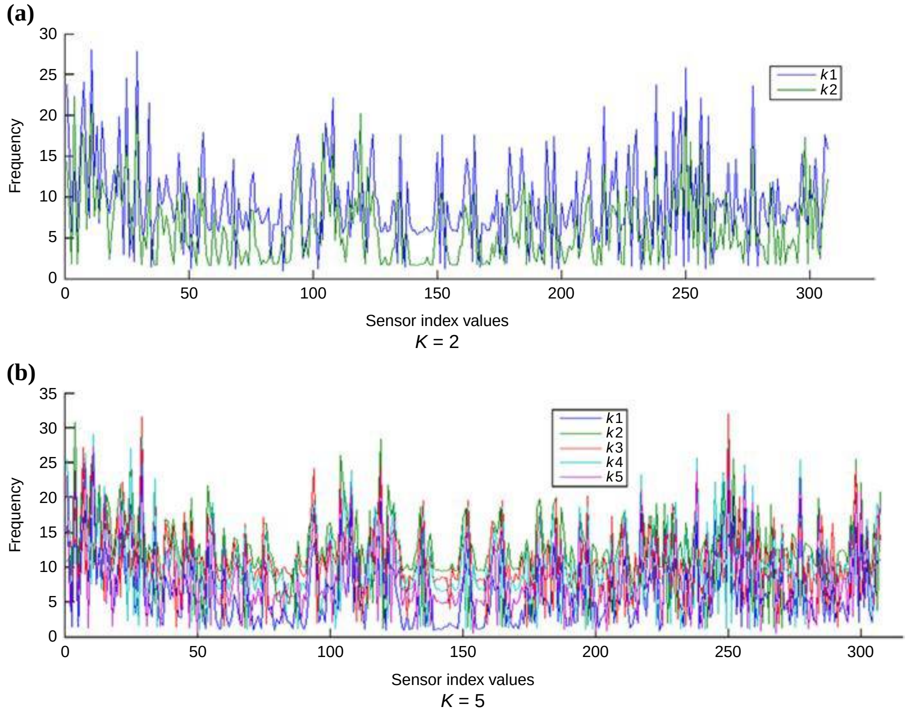


Figure 6. Sensor data clustering into three categorical patterns

events-to-cluster-centroid distances, and d includes all the distances between each event and cluster centroid.

For a better visualization of the sensor data behavioral pattern, we also analysis the same sample data in different techniques, which gives a succinct graphical representation of how well each event's lies within its cluster. As a_i be the average dissimilarity of e_i with all other events within the same cluster. However, distance measurements did not always reveal the better characteristics or behavior of the events to be in the same cluster. Therefore, a_i can be interpreted as how e_i assign to its cluster. The smaller the value is, the better the assignment that the average dissimilarity of point e_i to a cluster C as the average of the distance from e_i to the events point in C. Let b_i be the lowest average dissimilarity of e_i to any other cluster that e_i is not a member. The cluster with this lowest average dissimilarity is said to be the "neighboring cluster" of e_i since it is the next best-fit cluster for point e_i and it is defined as:

$$s_i = \frac{b_i - a_i}{max\{a_i, b_i\}} \tag{18}$$

Where the event similarities can be $s_i = \begin{cases} 1 - a_i/b_i & if \quad a_i < b_i \\ 0 & if \quad a_i = b_i \\ b_i/a_i - 1 & if \quad a_i > b_i \end{cases}$ as S_i is always between −1 and 1 $(-1 \leqslant s_i \leqslant 1)$.

Based on this computational analysis, the sample sensor data synthesized in the same iteration and class categories, which provide the result as showed Figure 7(a) and (b).

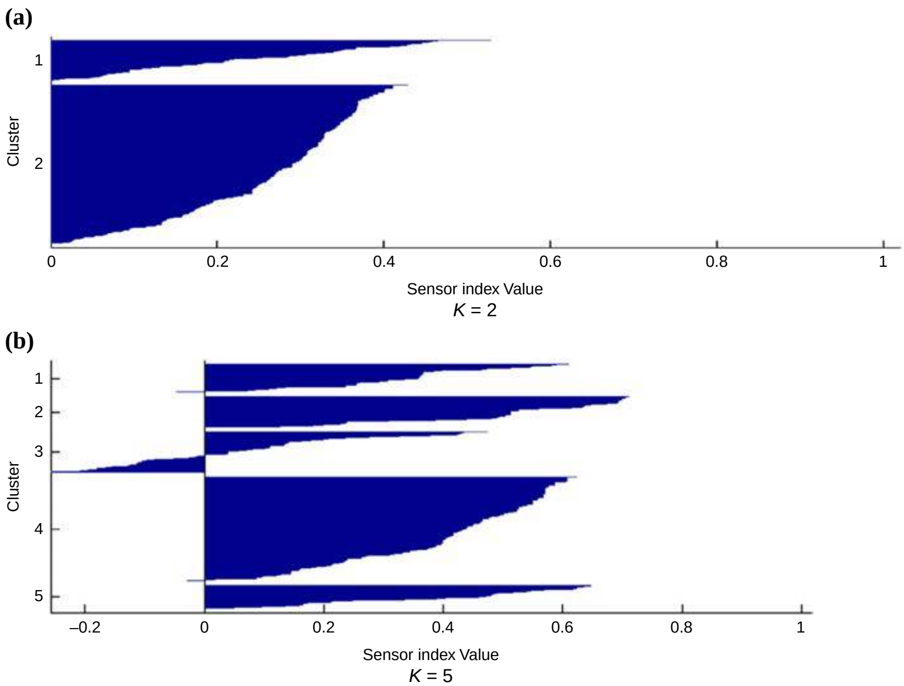


Figure 7. Alternative sensor data pattern clustering representations

The event patterns run from 2-5 different cluster values using the *k*-means routine from the Statistics Toolbox on the Sample taste the data matrix *X* from Table V.

As it is shown in Figure 7, the sample data behavioral pattern is more detectable as the number of clusters is increasing. Moreover, the data categories become more selective and give clear insights to interpret the real situations of SE to address its optimal performances and fault detections. For instance, in Figure 6(a), the events clustered behavioral patterns showed a similar tendency to Figure 6(b). The event patterns in Figure 7(b) against Figure 7(a) clearly visualizes data types, which provide more understanding of the behavioral patterns. It is vitally essential to support the decision-making processes by focusing on better or more probable events towards SE performance and fault detections. Furthermore, the sensor data clustering evaluated by the DTW-based distance measurements as of Equations (7) and (8), the outcomes visualized as Figure 8. The lesson of the outcomes is how the pattern is oscillating, which is pertinent to define the events that need to be analyzed as the domain contexts and users desire.

Sensor data are always complex and active, which might not completely store. Because of these facts and other data, behavioral pattern clustering cannot be prefaced to define the data situations of nor of data normal and abnormal pattern. Therefore, some outcast data can be explored as their time and location positions. The computational algorithm is defined as Equation 15 to validate the outcomes as a normal or abnormal event's behavioral pattern:

$$f(a, k_a) = \frac{1}{N} count(k_a) \tag{19}$$

where N is the total recorded activities or number of observed behavior counts, (k_a) is the number of events within the cluster k where the activity α resides. The minimum threshold is also computed and defined as Equation (16) to visualize the abnormal pattern behavior as it is shown on Figure 9. Therefore, the calculated value of θ is 0.99, which is less than the confidence or normal value (1):

$$f(a, k_a) < \theta \rightarrow \textit{abnormal behavior} \tag{20}$$

As it is shown in Figure 9, the abnormal behavior of sensor data recognized in three distinctive categories. These are 'no records or activities, high activities that are the normal or acceptable recognition and mixed that are acceptable or non-acceptable activity's recognition. However, in this visualization, the detected ABP needs further isolations and identification processes. The insignificant and less acceptable activity recognitions treated as an ABP, which are essential to investigate the performance and challenges of SE intelligent services. Therefore, the data clustered behavioral patterns need to be computed based on the data internal relation behaviors. Based on this analysis, the distance measurement value with correlations and event's weight deviation showed universe relations, which visualize the correlation matrix of the sensor data interactions as showed on Figure 10.

The correlation pattern model view of Figure 10 visualizing the isolated ABP as less correlation behavioral pattern, which is pertinent to a critical analysis of the dominant factors of the smart space in terms of its recorded data. The outcome is important to detect and interpret the challenges of SE performances and applications, which includes sensor device's locations, data management, preprocessing and other related issues to improve the performance of smart systems.

As the DM based Naïve Bayes process of event distributions and similarity measures (hidden relationships) of sensor data are also essential to analyze events performance towards ABPs as discussed in Sections 3.3 and 3.4. The computational outcomes are tabulated and represented as shown in Tables VI and VII naïve Bayes simple distribution and events similarity, respectively.

The Naïve Bayes simple distributions based graphic representations of events behavioral pattern clearly visualized as Figure 11(a)-(e), which is vitally significant to distinguish the ABPs. Based on these computational facts (the variations, between events, mean and standard deviation), the ABPs on Figure 11(a)-(e) are considerably isolated. Whereas for Figure 11(d), the factors are in between nor normal and abnormal patterns. Furthermore, events variations distributions also computed as Table VII.

The parameters S_1, S_2, S_n ,…, relative distance measurement is also important to gain a clear understanding of sensors data behavioral pattern devotions in relation to

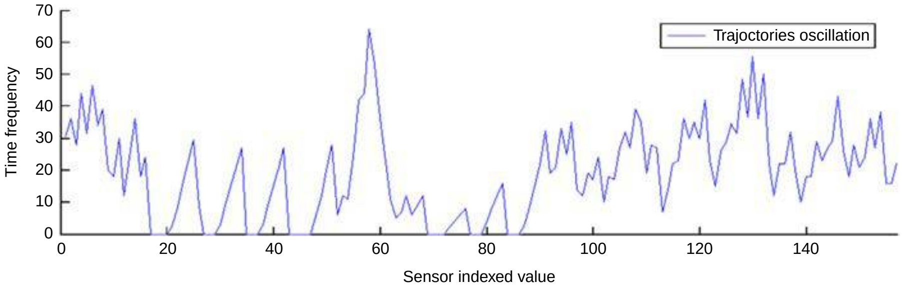


Figure 8.
Dynamic time warping based sensor data pattern behavior

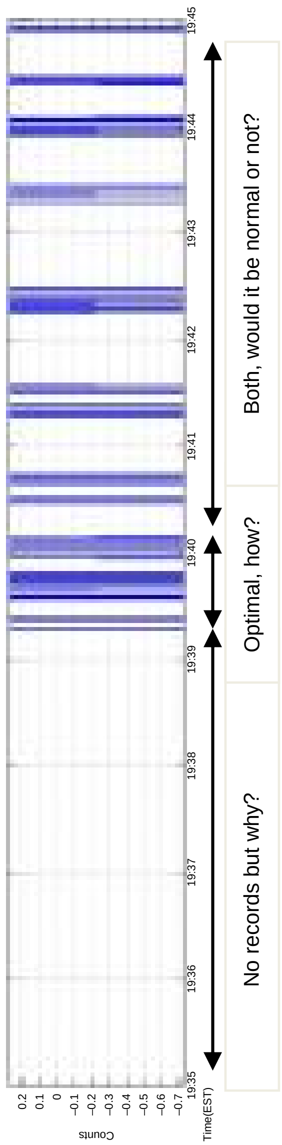


Figure 9.
Sensors data patterns across the given time

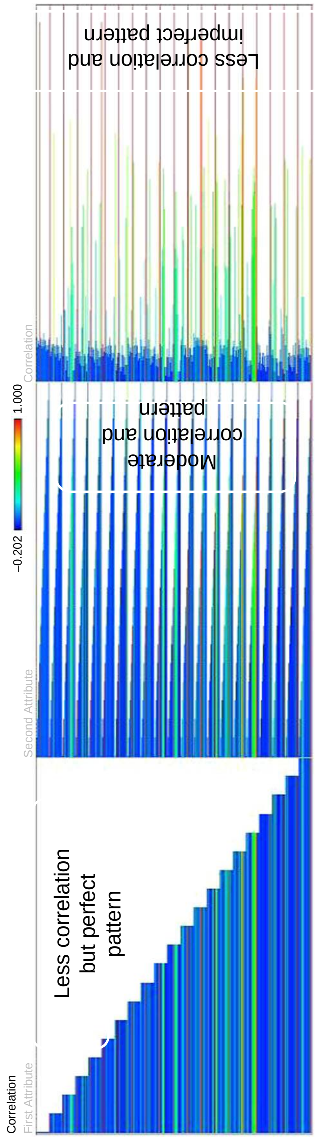


Figure 10.
SE sensor data correlation behavioral pattern view

sensors locations and data capturing capacities in the intelligent systems. The hidden interactions of data are a fundamental factor to discriminate the data behaviors fully. Therefore, HMM-based data similarities or distance variation model visualized as Figure 12. In an HMM-based time series clustering algorithm of sensor data, the distance measure is an indicator to select the proper patterns in a basic clustering process. However, HMM-based behavioral pattern performance analysis is complex and focussing on measuring to discriminate data similarities than behavioral deviations or fault diagnosis, which is well adopting in speech recognitions, image segmentations, and others. On the other hand, DM clustering and nearest neighbors techniques are dynamic and adaptable, and also a systematic approach to augment applications for sensor data behavioral pattern detections. Based on these facts, a DM technique for sensor data BPM is an advanced and scalable application in the engineering fields, which can be an inference to many other scientific works. It has powerful and capable performances to define time series streaming data similarity or deviation measurements.

V. Conclusion and future directions

In this paper, we discussed the innovative topic of critical analysis of sensor data behavioral patterns, which is pertinent to SE performance optimizations and fault diagnosis. In the first section, we discussed SE or intelligent systems in general in order to understand the applications the smart technologies in relations to sensor data deluge. The large-scale sensor data need to be analyzed to define the trajectories' behavioral patterns, which is the theme of this study. The technique of SDM focussed on in the paper is to explore or mine sensor data behavioral patterns in a smart place to detect the abnormal behavioral patterns for fault analysis and SE performance optimizations in general. In Section III that followed the relative work discussed in details the sensor data behavioral patterns in terms of sensor data ecosystems, characteristics, pattern development and descriptions. Besides to these, the technique of SDM of clustering and distance measurements based behavioral patterns explored, pattern-clustering algorithms proposed, and various distances measurements-based techniques synthesized. In Section IV, the proposed method performance, their outcomes and the need for the approach well analyzed and evaluated using real sensor data. The experiment was conducted to elucidate sensor data behavioral patterns that

Attributes (sensors index)	*S*1		*S*2		*S*3		*S*4		*S*5		…
Parameters	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	…
Predicted pattern behavior											
Normal (y)	0	0.001	0.154	0.555	0.231	0.599	0.231	0.599	0.308	0.751	…
Abnormal (n)	0.250	0.707	0.750	2.121	0.750	1.488	0.125	0.354	0	0.001	…

Table VI. Naïve Bayes simple distributions

1st Att. (sensors index)	*S*1	*S*1 …	*S*2	*S*2 …	*S*3	*S*3 …	*S*4	*S*4 …	*S*5	*S*5 …	…
2nd Att. (sensors index)	*S*2	*S*3	*S*1	*S*3	*S*1	*S*2	*S*1	*S*2	*S*1	*S*2	…
Distance value	6.708	6.245	8.602	6.782	6	6.164	3.742	6.782	6.928	3.742	…

Table VII. Similarity measurement of sensor data

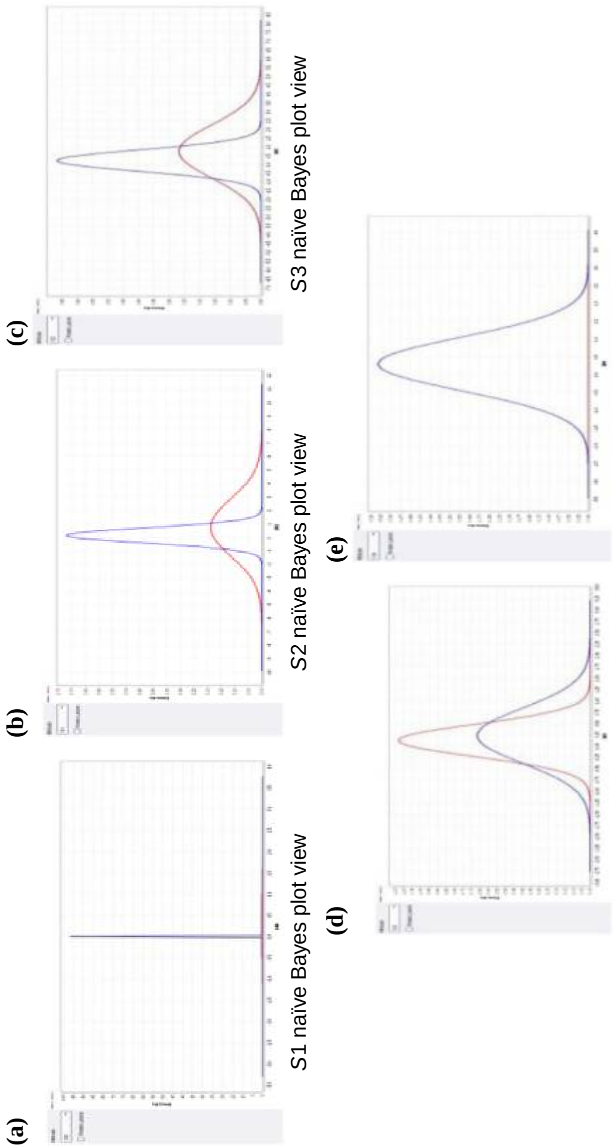


Figure 11.
Naïve Bayes event distributions graphic representation

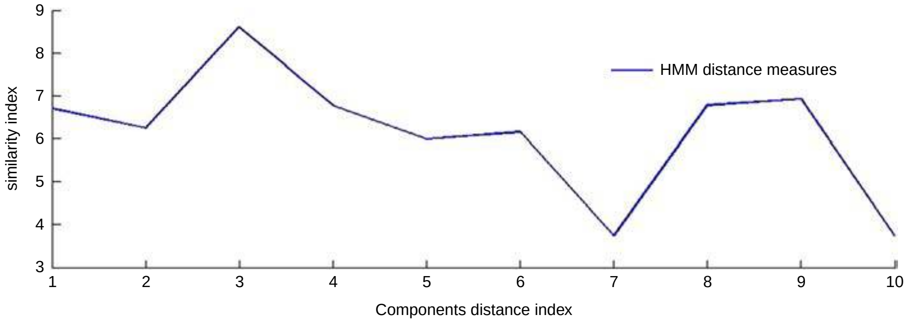


Figure 12.
HMM-based sensor data behavioral pattern model

semantic enrichment can enhance ABPs detections. The results presented in graphs and interpreted for users, and for further development and applications of sensor technologies. Thus, the SDM-based sensor data critical analysis augmented as to sensor data behavioral pattern mining toward ABPs detections, which helps to isolate, identify and accommodate smart space faults as the domain contexts.

In the past studies, it was demonstrated that only a limited amount of information extract from raw sensor data. In this study semantics, sequential events BPM as to ABPs detections provide novel ideas to extract knowledge implicit, which SDM could scale up and play a great role in the new era of big sensor data. Furthermore, the study has some limitations that can be venues for future research. This study did not consider the causative factors of the ABPs that are how the faults or changes happened, which can provide more specific insights into the root challenges of the sensor functionality. It needs to extend how to control and reconfigure the abnormal behavior. Sensor data abnormal behavior identification in relation to real-time SE risk assessments is vitally pertinent for advanced and large-scale intelligent systems.

Reference

Ahmad, L. *et al.* (2011), "Smart homes for the elderly dementia sufferers: identification and prediction of abnormal behavior", *J Ambient Intell Human Compute*, Vol. 3 No. 3, pp. 205-218.

Ahmed, E. and Nada, E.-G. (2013), "Trajectory data mining: integrating semantics", *Journal of Enterprise Information Management*, Vol. 26 No. 5, pp. 516-535.

Ariel, K. *et al.* (2012), "The big data bootstrap", *Proceedings of the 29th International Conference on Machine Learning*, arxiv.1206-6415v1, pp. 1759 -1766.

Bhatia, M.P.S. and Deepika, K. (2013), "Experimental Study of data clustering using k-means and modified algorithms," *International Journal of Data Mining & Knowledge Management Process*, Vol. 3 No. 3, pp. 17-30.

Brian, H.R. and C.J.H., M. (2007), "Smart systems and environments", *Kybernetes*, Vol. 36 Nos 3/4, pp. 456-464.

CASAS (2010), "Smart home publicly shared datasets", available at: http://ailab.wsu.edu/casas/datasets.html (accessed 6 November 2012).

Charu, C.A. (Ed.) (2013), *Managing and Mining Sensor Data*, Springer, New York, NY.

Chen, L. *et al.* (2005), "Robust and fast similarity search for moving object trajectories", *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 491-502, doi:10:1146/1066157-1066213.

Debraj, D. *et al.* (2012), *ActiSen: Activity-Aware Sensor Network in Smart Environments*, Elsevier B.V., Pervasive and Mobile Computing, Vol. 8 No. 5, pp. 730-750.

Dhaval, P. *et al.* (2009), "Analyzing abnormal events from spatio-temporal trajectories", IEEE *International Conference on Data Mining Workshops*, pp. 616-621, 978-0-7695-3902-7/09, doi10.1109/ICDMW.45.

Diane, J.C. and Lawrence, B.H. (2011), "Sensor selection to support practical use of health-monitoring smart environments", *WIREs Data Mining and Knowledge Discovery*, Vol. 1 No. 4, pp. 339-351.

Dong, Z. *et al.* (2005), "Semi-supervised adapted HMMs for unusual event detection", *Proceedings of the 2005 IEEE Computer Society Conference on CVPR*, No. 10636919/ 05, pp. 611-618v1, doi:10:1109/CVPR.2005.316.

Elad, Y.-T. (2004), "An introduction to pattern classification", *Machine Learning 2003, LNAI 3176*, Springer-Verlag Berlin, Heidelberg, pp. 1-20, doi:10:1007/978-3-540-26550-9-1.

Erfaneh, A., Mohammad, A. and Heidari, J. *et al.* (2012), "The role of smart home in smart real estate", *Journal of European Real Estate Research*, Vol. 5 No. 2 pp. 156-170.

Fan, J. *et al.* (2009), "A dynamic hierarchical clustering method for trajectory-based unusual video event detection", *IEEE Transactions on Image Processing*, Vol. 18 No. 4, pp. 907-913.

Fan, J. *et al.* (2011), "Anomalous video event detection using spatiotemporal context", *Elsevier, Computer Vision and Image Understanding*, Vol. 115 No. 3, pp. 323-333.

Fernando, R.-I. *et al.* (2005), "A neural network agent-based approach to activity detection in ami environments", IEE International Workshop, Intelligent Environments, Colchester, 28-29 June, ISSN 0537-9989.

Jaakko, S. *et al.* (2008), "Gaussian process person identifier based on simple floor sensors", EuroSSC 2008, LNCS 5279, Springer-Verlag Berlin Heidelberg, pp. 55-68, doi:10:1007/978-3-540-88793-5-5.

Jae-Gil, L. *et al.* (2007), "Trajectory clustering: a partition-and-group framework", SIGMOD'07, ACM 978-1-59593-686-8/07/0006, doi: 10:1146/1247801247546, pp. 593-604.

Jan, O.B. (2001), "A pattern approach to interaction design", *AI & Soc*, Vol. 15, Springer – Verlag, London, pp. 359-376, ISBN-10:0471498289, Ed1.

Jianping, Z. *et al.* (2010), "A new distance measure for hidden Markov models," *Expert Systems with Applications*, Vol. 37 No. 2, pp. 1550-1555.

Joëlle, C. *et al.* (2005), "Context is key: communications of the ACM", Vol. 48 No. 3, pp. 49-53.

Juan, C.A. *et al.* (2010), "Ambient intelligence and smart environments: a state of the art", Springer Science + Business Media, Llc, Part 1, doi:10.1007/978-0-387-93808-0_1, pp. 3-31.

Karthika, L. and Sumathi, A.C. (2012), "Automatic discovery and recognition of activities in smart environment", *International Journal of Modern Engineering Research*, Vol. 2 No. 1, pp. 353-358.

Mu-Yen, C. and Edwin, D.L. (2013), "Smart space technology innovations", *Library Hi Tech*, Vol. 31 No. 2, pp. 197-200.

Oren, B. and Michal, I. (2005), "Detecting Irregularities in Images and Video", *Proceedings of the Tenth IEEE International Conference on Computer Vision*, No. 15505499/ 05.

Parisa, R. and Diane, J.C. (2010), "Mining Sensor Streams for discovering human activity patterns over time", *IEEE International Conference on Data Mining*, 1550-4786/10, doi10.1109/ICDM. 40.

Parisa, R. and Diane, J.C. (2011), "Activity knowledge transfer in smart environments", *Elsevier, Pervasive and Mobile Computing*, Vol. 7 No. 2011, pp. 331-343.

Parisa, R. *et al.* (2011), "Discovering activities to recognize and track in a smart environment", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 23 No. 4.

Paul, C. and Richard, M. (2001), "An adaptive approach to indexing pervasive data", ACM 1-58113-412-6/01/05.

Peng, L. and Elia, E.-D. *et al.* (2007), "Applying data mining algorithms to inpatient dataset with missing values", *Journal of Enterprise Information Management*, Vol. 21 No. 1 pp. 81-92.

Piciarelli, C. and Foresti, G.L. (2006), "On-line trajectory clustering for anomalous events detection", *Elsevier, Pattern Recognition Letters*, Vol. 27 No. 15, pp. 1835-1842.

Rainer, S. *et al.* (2008), "Tracking identities and attention in smart environments – contributions and progress in the CHIL project", *IEEE, Automatic Face and Gesture Recognition, 8th International Conference*, pp. 1-8, 978-1-4244-2154-1/08/c, doi: 1109/AFGR2008.4813322.

Ramakrishnan, S. and Rakesh, A. (2013), "Mining sequential patterns: generalizations and performance improvements", *Proceeding, the 5th International Conference on Extending Database Technology*, pp. 3-17.

Stephen, J.P. and Ernest, D.D.S. (1997), "Smart sensors for industrial applications", *Sensor Review*, Vol. 17 No. 3, pp. 217-222.

Szewcyzk, S. *et al.* (2009), "Annotating smart environment sensor data for activity Learning Technology and HealthCare", Vol. 17 No. 3, pp. 161-169.

T^am, H. and Bernt, S. (2006), "Unsupervised discovery of structure in activity data using multiple eigenspaces", Springer-Verlag, Berlin Heidelberg, LNCS 3987, pp. 151-167.

Tao, G. *et al.* (2009), "epSICAR: an emerging patterns based approach to sequential, interleaved, and concurrent activity recognition", *Proceedings of the IEEE International Conference on PCC*, 978-1-4244-3304-9/09/$25.00, pp. 1-9, doi: 1109/PERCOM2009.4912776.

Tao, W. (2012), "An efficient algorithm for mining condensed sequential pattern basis", *Kybernetes*, Vol. 41 No. 9, pp. 1289-1296.

Thomas, B.M. *et al.* (2006), "A survey of advances in vision-based human motion capture and analysis", *Elsevier, Computer Vision and Image Understanding*, Vol. 104 Nos 2-3, pp. 90-126.

Wiliem, A. *et al.* (2008), *Detecting Uncommon Trajectories in Digital Image Computing: Techniques and Applications (DICTA)*, Canberra, pp. 1-3, doi: 10:1109/DICTA2008.45, pp. 398-404.

Yan, H. *et al.* (2008), "A framework for mining sequential patterns from spatio-temporal event data sets", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20 No. 4, pp. 433-448.

Yiu, C.Y. and Yau, Y. (2006), "A learning model of intelligent home", *Facilities*, Vol. 24 Nos 9/10, pp. 365-375.

Zhonghong, W. (2008), "Smart Spaces: creating new instructional space with smart classroom technology", *New Library World*, Vol. 109 Nos 3/4, pp. 150-165.

Further reading

Jianhua, M. and Laurence, T.Y. *et al.* (2005), "Towards a smart world and ubiquitous intelligence: a walkthrough from smart things to smart hyperspaces and UbicKids", *International Journal of Pervasive Computing and Communications*, Vol. 1 No. 1, pp. 53-68.

About the authors

Dr Gebeyehu Belay Gebremeskel is a Postdoctoral Research Fellow in Chongqing University, China. He received his PhD from the Chongqing University, since July 2013, Master's Degree from the London South Bank University, UK and BSc Degree from the Alemaya University, Ethiopia. Dr Gebeyehu gained solid experience in research and teaching in different positions and institutes. He also engaged and contributed fundamental professional skills in various IT projects and research activities. In his PhD and Master's studying program, he published more than 25 academic papers, which includes journals, conference papers, book chapters and other technical reports in his research area and computing field in general. His research interest includes Data Mining, Business

Intelligence, Big Data, Machine Learning, Agent Technologies, Artificial Intelligence, Data cloud, and others in the field of Computer Science. Dr Gebeyehu Belay Gebremeskel is the corresponding author and can be contacted at: ge.be09@yahoo.com

Professor Chai Yi is currently the Vis Director of College of Automation in the Chongqing University. He holds a PhD Degree major in Control theory and Engineering and Master's Degree major in Industry Automation from the Chongqing University, and his first degree from the National University of Defences Technology major in Electronic Engineering, China. Professor Chai is a distinguished Full Professor in the Chongqing University. He is a Principal Postgraduate Courses' Lecturer and PhD Tutor, honourable academician and Researcher in his field and in the college in general. Because of his deep academic scholastic, he innovated and running more than 40 national projects. Professor Chai is the famous and known Academician, and in his successful achievement and patented work, he received more than nine national awards. He is also a Member and the Head of many professional and technical associations in the national level. He is therefore, highly respectful and honoured by his colleagues and students because of his immense and exemplary efforts and wishes to deliver applicable and scalable scientific techniques, methodologies, knowledge and experience to his followers. He published many academic and high standard journals, books, and other technical reports in both English and Chinese. His major research direction is information processing, fusion and control, industrial process control theory and technology, intelligent systems theory and applications, and others in the field of automation.

Professor Chengliang Wang received BA in Mechatronics, MA in Precision instruments and machinery and the PhD in Control Theory and Control Engineering from the Chongqing University. Currently, he is a Professor been long engaged in artificial intelligence theory and application, teaching and research in computer application technology. He is a Principal Lecturer and Scientist at the College of Computer Science, Postgraduate Students Supervisor and Advanced Fellow of the Chinese Computer Federation. His research is mainly in the field of complex system processing, data mining, image processing, artificial intelligence theory and application, and so on. He has published more than 20 academic articles indexed in key journals, which are indexed by SCI or EI. Professor Chengliang holds two inventive patents. He also innovated and actively participating more than nine national projects and seven patent technologies.

Zhongshi He is a Professor and the Vice Chair of the College of Computer Science in the Chongqing University. He is also a distinguish Lecturer and Scientist in the college, and a Visiting Professor in few other high standard national and international universities. He received a BSc, Masters in Applied Mathematics and PhD degree in Computer Science from Chongqing University, and Post-doctoral fellowship from Witwatersrand University in South Africa. His research and teaching activities have focussed on Machine Learning and Artificial Intelligence issues. He is Masters, PhD and Post-doctoral students' supervisors. In his success, solid skills and interest to share his cumulative knowledge and experience, highly respected and honoured in the college and in the University as a whole. So that he is one of the tops honoured Senior Instructors. His research themes include workstation design based research projects in Machine Learning and Artificial Intelligence. Therefore, his research interests center on improving the understanding, design, and performance of machine learning, image processing, artificial intelligent, data mining and bioinformatics focussing on its applications and performance optimizations. In the Machine Learning arena, he has worked on many practical and applicable researches. He has explored complex and diversified research issues in computer science, which has made numerous contributions to the field. Most of his publications and conference presentations deal with organizational issues on various real-world research issues. He has given numerous invited talks and tutorials in the other universities too. Professor He is published many high standard journals, conference papers, and other technical reports. Besides his administrative duties, Professor is a Faculty Member at the College of Computer Science and responsible for many executive and academic tasks.

This article has been cited by:

1. Gebeyehu Belay Gebremeskel State Key Laboratory of Power Transmission Equipment, System Security and New Technology, College of Automation, Chongqing University, Chongqing, China Chai Yi State Key Laboratory of Power Transmission Equipment, System Security and New Technology, College of Automation, Chongqing University, Chongqing, China Zhongshi He College of Computer Science, Chongqing University, Chongqing, China Dawit Haile School of Engineering, Science and Technology, Virginia State University, Petersburg, Virginia, USA . 2016. Combined data mining techniques based patient data outlier detection for healthcare safety. *International Journal of Intelligent Computing and Cybernetics* **9**:1, 42-68. [Abstract] [Full Text] [PDF]