Human Behaviour Analysis Using Data Collected from Mobile Devices

[1]Muhammad Awais Azam, [1]Jonathan Loo, [1]Sardar Kashif Ashraf Khan, [2]Muhammad Adeel, [3]Waleed Ejaz
[1]School of Engineering and Information sciences, Middlesex University, London, UK
[2]School of Electronic Engineering and Computer Science, Queen Mary University, London, UK
[3]Department of Information and Communication Engineering, Sejong University, Republic of Korea
*Corresponding Author {m.azam@mdx.ac.uk}

Abstract- **Human behaviours are multifarious and myriad in nature. It is a challenging task to envisage and learn the human behaviour from daily routine activities. The profusion of wireless enabled mobile devices in daily life routine and advancement in pervasive computing has opened new horizons to analyse and model the contextual information. The aim of this research work is to infer the behaviour of low entropy mobile people using contextual data collected from mobile devices such as GSM location patterns (cell tower ID data) and Bluetooth proximity data. Both the GSM and Bluetooth data itself do not reveal much information about the behaviour of the users. Therefore, the challenge is to find out whether such data can infer human behaviour to understand and aid the unusual activities and routines of low entropy people such as elderly people and early stages of dementia patients. In this paper, a framework is created to analyse the contextual data for behaviour detection. There are four different steps in this framework to achieve the objective of the research work. In the first step, the contextual data is first classifies into different locations to obtain the movement patterns of the users. In the second and third step, a probability matrix and training data is obtained respectively, depending upon the user's movement on daily and hourly basis. In the fourth step, a decision engine i.e. Neural Network (NN) and Decision Trees (DT) is used to detect the behaviour of the low entropy user. Results have shown that cell tower ID data gives behaviour of the user on high level scale for example movement patterns in GSM cells that does not help to identify any lower level activities such as attending the lecture, traveling in a bus. Whereas, Bluetooth data gives us more information about the lower level activities depending on the social relations and close proximity of other users.**

Keywords – Behaviour, Cell Tower ID, Bluetooth Proximity, Neural Networks, Jaccard Index, Decision Trees

I. INTRODUCTION

Detection and prediction of human behaviour from daily life activities is a challenging task. People can have both regular and varying daily life routines that make it a burning topic nowadays in social research circles. Modelling human behaviour such as individual routines from proximity data and social relations with gathered data of daily life activity patterns is an emerging realm in Ubiquitous Computing. Computers are becoming more and more pervasive and are embedded in everyday objects, such as cameras, music players, cars, clothing etc. There can be different sensing devices e.g., Radio Frequency Identification (RFID), motion sensors, GPS enabled tracking devices, and other context aware devices that can be used for real time proximity detection and daily life data gathering purposes. In particular, devices such as mobile phones provide a rich platform for various forms of data gathering by using its integrated sensors such as Bluetooth ID, digital camera, microphones and GPS transceivers. These sensors can give an individual's location, movement and proximity information for the whole period of cell phone usage. Specifically, Bluetooth radios are frequently incorporated into mobile devices [2].

This new generation of "smart devices" has created new ways to utilise the capability of computers and enhanced the area of Ubiquitous Computing by providing rich and detailed information about the context of the user. Context-aware computing, which is part of Ubiquitous Computing, uses sensors either in the environment or carried / worn by the users to extract and interpret the user's context, for example what resources are available, who is in close user's proximity. This contextual information can help to recognise different tasks and activities perform by the user.

Different researchers have worked on routine and activity classification using mobile phone data [2] [3] [13] [14]. They have tried to analyse the social relationships and daily life routine patterns of individuals using their cell phone data. They classified the cell tower ID data into different locations such as Home, Office, Elsewhere and No-signals and analyses the movement patterns. They have also used Bluetooth proximity data to differentiate between weekday or weekend activities. In our research work, we want to go a step further in behaviour analysis. As cell tower ID information can only give patterns of movement and location information. It cannot tell us about low level activities. For example, cell tower ID data can tell whether the user is at home or in office/campus, but it cannot tell in which activity such as attending the lecture or sitting in cafeteria, user is participating. On the other hand, proximity data can give information about the people and other devices that are in close vicinity of the user but it does not tell exact information about the user's location. If this proximity data can be classified into different locations, then proximity information can provide a good idea about the nature of activity that user is performing and this will help in analysing and understanding an individual's behaviour.

The aim and focus of this research work is on the detection of behaviour of the people who live low entropy lives that means they follow somewhat regular routines and exhibit less change in their behaviours as discussed in [3]. According to [3], if the user in his daily life, repeat the activities and routines with less change, it will be known as 'low entropy' behaviour. While a more change in daily routine patterns is considered as 'high entropy' behaviour. For example, a working person who follows the routine of going to the office and coming back home every day using the same means of the transport, or an elderly person with regular routines [4] (e.g., an early stage of dementia patient) can be the examples of the people with more regular routines and hence less change in the behaviour. The motivation of this research work is to help elderly people and early stages of dementia patients to live their lives more independently by understanding their behaviour from wireless proximity data.

This work is an extension of [1] and [5], in which repeated patterns and behaviour of an individual was detected by using n-gram technique and considering only Bluetooth proximity data and the behaviour was detected by using only NN. The research work in [5] proves the concept that daily life traces of Bluetooth proximity data of a low entropy individual can give us enough repeated patterns in the data that can be further used for activity or behaviour detection. In [1], the unusual routines in the daily life of the user were detected by using the NN only.

In this research paper, we have used two different types of contextual data (GSM cell tower ID and wireless proximity data), that are rather collected independently, to analyse the behaviour of low entropy mobile people. Wireless proximity data that is used in this research work is of Bluetooth. The data set used in this paper is the reality mining dataset [3] collected at MIT for the year 2004-2005. Nokia 6600 cell phones were used to record the data of 100 users over the duration of 9 months. Different types of information were collected including phone status i.e. whether it is in use or charging or off, ID's of Bluetooth proximate devices, usage of mobile applications, cell tower ID data, call and SMS logs

The rest of the paper is as follows: Section-II contains related work on unusual activity detection and usage of Bluetooth as a sensing device. Section-III and Section-IV discusses the research objectives and the behaviour analyses framework respectively. Section-V discusses the behaviour analysis results using cell tower ID data and Section-VI contains the results of behaviour detection using Bluetooth proximity data. Summary of the work and notes on the direction planned for the future work is in Section-7.

II. RELATED WORK

Detection of abnormality in human behaviour is very intricate and has been a challenging task in the past. Though, recent advancements in information technology had made it quite simpler. In last few years a lot of efforts have been made to observe the abnormal routines and daily life patterns of an individual [6] [7]. In [6], the author has presented a framework for the detection of unusual human behaviour inside an intelligent house. The author used motion sensors to detect the activities and unusual human behaviour patterns based on Markov Chain. Vector quantization is employed to reduce the sensor states and the transition between states is represented by probabilistic model. The above mentioned technique detects the unusual human behaviour either by computing the distance between the state transition probabilities or by the likelihood of human action. The distance between the state transition probabilities was calculated by using either Kullback-Leiber distance or Euclidian distance. Limitation of this work is that they only consider the indoor activities that can only happen inside the home. To analyse human behaviours and activities, some authors have also used devices other than motion sensors such as, accelerometers, digital cameras and microphones.

In literature some techniques has also been presented to analyse the accumulative behaviour of multiple individuals instead of one single individual. For example, in [8] the author proposed a framework based on identification of close proximity social behaviours. This work also focused on the movements inside a building. Similarly other multiple individual behaviour detection schemes such as group actions in meetings [9] and audio visual perception of a lecture in smart environment [10] are presented. However, majority of work in above mentioned studies have focused on indoor environment; as it is based on sensing devices which have several limitations such as short range of detection, less battery power and storage, or may not be very common that every person can use it without extra hardware, which is not feasible for outdoor environment.

The enormous penetration ability of Bluetooth technology have made it more suitable candidate to be used as a personal identifier. This capability can be exploited by using the mobile phone having Bluetooth technology as a sensing device. Nowadays the mobile phone is an indispensable part of our society with many types of embedded sensors. These sensors have been used in many worth mentioning applications such as social proximity sensing [11] [12], social behavioural modelling and routine classification [2] [3] [13] [14] and movement prediction [15] [16]. The significance of aforementioned studies is that these techniques have focused on how to recognize an individual's behavioural patterns and social routines but no one of them has classified the Bluetooth proximity data into different locations and predicts the behaviour by using the machine learning techniques.

In [13] and [14], researchers have presented a framework for daily life activity recognition based on the user's location and group affiliation. They used Author Topic Model (ATM) and hierarchical Bayesian topic models like Latent Dirichlet Analysis (LDA) for routine classification. The routines they classified are whether it is a weekday or a weekend depending upon the location of the user or the proximity information and whether the experimental subject is an engineering student or a business student. The proximity data is only classified depending upon the number of proximate devices. There classification of proximity data does not give any information about the location of the user.

In [15] and [16], NN are used to detect and predict user movement based only on cell tower IDs. They utilised the probability of user being at different locations. Our work is similar in one aspect with their work and that is; we have also utilized the probabilities of user being in different locations. Difference between our work and the work presented in [15] and [16] is that we have used real time data for our experiments and have used both cell tower ID and Bluetooth proximity data. In [17], researchers proposed a relaxing minimum description length (MDL) principle in order to build compatible decision trees that are suitable for novel behaviour detection. This relaxing MDL principle is to exploit additional tests/features in order to discriminate between normal and abnormal behaviours.

In [7], researchers detect abnormal event in solitary elder's daily life by mining the related data gained by sensors. They employ the association rules finding algorithm with time cluster to analyse the elder's activities. In first step, they cluster each item of elder activity with time and then in the second step, all frequent item sets were found and strong association rules were created. Researchers in [18] work on the recognition of abnormal activities based on the Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM). They incorporate a Fisher Kernel into the One-Class Support Vector Machine (OCSVM) to filter out the most likely normal activities. Then from those normal activities, they derive a model to detect abnormal activities and tried to reduce false positives. In [19], researchers have presented a model for abnormal behaviour detection. That model considers user's location based on the cell tower ID and used Dynamic Bayesian Networks (DBN) to predict user's location. They proposed an X-Factor model, which is a DBN with a hidden variable. User's location according to this model not only depends on the hour of the day and day of the week but also this latent variable that represents the abnormal behaviour.

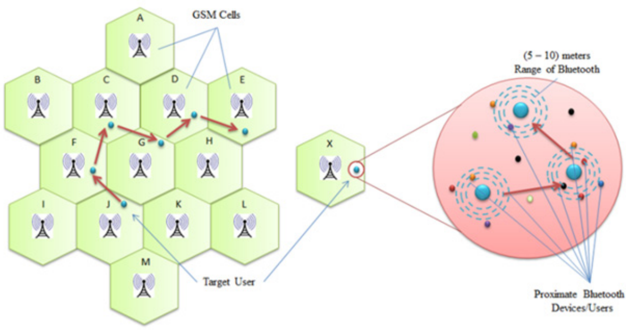
Most of the researchers as discussed above have focussed on routines and activity detection in closed and indoor environments and have used short range sensors that can work only in very close vicinity and have short battery life. This type of sensors cannot be used for outdoor environment. Our research work is not constrained of short range sensors and short battery lives. We explored the concept of mobile phone as a sensing device. Many other researchers as discussed above have also used mobile phones to get the proximity data and user's location from cell tower ID information. To the best of our knowledge, no one has classified the Bluetooth proximity data into different locations and obtained the user's movement patterns. In this paper, we address this concept and analysed the behaviour from cell tower ID and Bluetooth proximity data and found out that Bluetooth proximity data alone can be used to detect the behaviour of the low entropy mobile user. Results have shown that patterns in wireless proximity data can give us enough information about the routines of the user and unlikely cell tower ID data that can only give indications of user movement patterns at different locations, it can also give information about user's activities while staying at one particular location which is not possible to get from cell tower ID data.

III. RESEARCH OBJECTIVES

The primary aim of this research work is to find any anomalies in the behavioural patterns or routine activities of low entropy mobile people in order to aid in the detection of any unusual behaviours in elderly people or patients such as early stages of dementia patients. First objective is to utilize the contextual data (such as, cell tower ID and Bluetooth proximity data) available around us that can be obtained through different sensing devices, especially mobile phones, for behaviour detection. A framework is designed to analyse the behaviour of the low entropy users by using this contextual data.

The nature of Cell tower ID and Bluetooth proximity data is different from one another. Figure-1(a) shows the movement of a user in between different GSM cell towers. When a user is in the range of any GSM cell tower, ID of the cell tower is detected. This cell tower ID data only gives information about the user's movement in broad overview and cannot tell what type of activities user is performing within the range of detected cell towers. For example in Figure-1(a), user was in cell 'J', then moved to the cells 'F', 'C', 'G', 'D' and 'E' respectively. This information can only tell about the user's movement patterns and cannot give any idea about the activities that user is performing while at these locations. The purpose is to utilise this cell tower ID data to analyse behaviour of low entropy mobile people from the 'location data'. In order to detect the behaviour, two different machine learning algorithms have been used in the framework. The detection accuracy of both algorithms is also studied.

On the other hand, Figure-1(b) shows the detection of Bluetooth proximate devices. Cell tower ID data only can give user's location information, which in this case is cell 'X', whereas Bluetooth proximity data gives information about the people and other Bluetooth devices that are within the range of user's Bluetooth mobile device. Social relationship and group activities can be detected with this proximity data which is not possible to detect from the cell tower ID data. A weakness of Bluetooth proximity data is that it does not give any direct information about the location of the user. Location information is important to know in order to analyse the behaviour from daily routines and activities of the low entropy users. To obtain the location information from the Bluetooth proximity data is a challenging task and is also an objective of this research work. To get the location information from Bluetooth proximity data, we classify the Bluetooth detected devices into different groups that belong to locations such as Home, office and inferred the location of the user depending upon the detected devices. Another objective is to find out whether only Bluetooth proximity data can be used for behaviour and activity analyses and whether it can add more information about activities and daily routines of the user if we consider both cell tower ID and Bluetooth proximity data together.



(a) (b)

Figure 1. Scenario of GSM Cell Tower and Bluetooth Proximate Devices Detection

IV. BEHAVIOUR ANALYSIS FRAMEWORK

Figure-2 shows the overall framework which is going to be used to analyse the behaviour of low entropy mobile people from cell tower ID and Bluetooth proximity data. As aforementioned, real time traces of GSM cell tower ID and Bluetooth proximity data of low entropy people used for this research work is obtained from the Reality Mining dataset. There are four steps in this framework that are followed to achieve the objective.

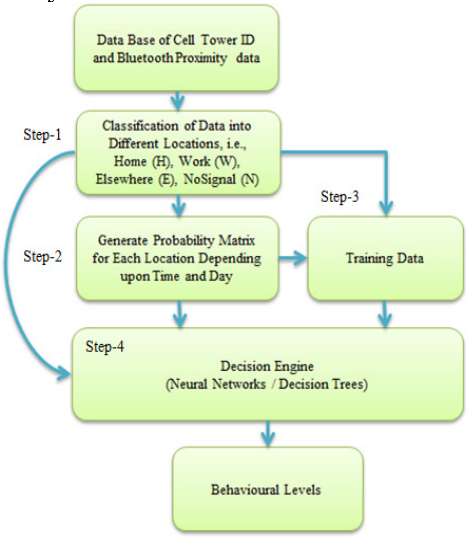


Figure 2. Behaviour Analysis Framework

Step-1 is to classify the cell tower ID and Bluetooth proximity data into different locations to find the activity and routine patterns of the user. The cell tower ID data which is obtained from the Reality Mining dataset is already classified into four different locations; i.e., Home (H), Work (W), Elsewhere (E) and NoSignal (N); whereas Bluetooth proximity data is in the form of list of detected proximate devices by the target user. Classification of Bluetooth proximity data into different locations is a great challenge and the classification procedure of Bluetooth proximate devices into different locations is explained in detail in Section-6.

Step-2 is to obtain a probability matrix predicting the location conditional on the hour of the day and day of the week from the classified information. This means, every entry of this matrix depends upon the specific hour of the day and whether it was a week day or a week end. Figure-3 shows the structure of the probability matrix for H, W, E and N for all twenty four hours. Each row in this matrix shows the hour of the day and each column shows the probability of H, W, E and N for that hour of the day. Depending upon this calculated probability; behaviour is divided into four different levels, shown in Table 1. Every entry of the probability matrix depends upon the specific hour of the day and whether it is a weekday or a weekend.

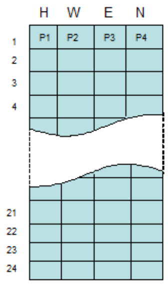


Figure 3. Probability Matrix for H, W, E and N

TABLE 1. BEHAVIOURAL LEVELS

Probability	Behaviour
0 < p < 0.25	Abnormal
0.25 < p < 0.5	Low Abnormal
0.5 < p < 0.75	Average Normal
0.75 < p <1	Normal

Step-3 is to utilise the probability matrix obtained in the setp-2 for preparing training data for the decision engine that is used for the detection of level of abnormality in the user's behaviour.

In step-4, the decision engine will use the training data obtained in the step-3, the probability matrix from step-2 and the classified data from setp-1 with a machine learning algorithm (NN or DT) to detect the behaviour of the user that deviates from the normal routines.

Next section discusses the behaviour analysis from cell tower ID data by using the above mentioned framework.

V. BEHAVIOUR ANALYSIS FROM CELL TOWER ID DATA

This section uses only the cell tower ID data of a low entropy user obtained from the reality mining dataset with the entropy level 23.06, calculated by using the Shannon's entropy equation shown in Equation-1, to find any anomalies in the daily life routines and behaviour of the user.

$$H(x) = -\sum_{i=1}^{n} p(i) \log_2 p(i) \qquad (1)$$

Cell tower ID gives information about the user's location and movement patterns. Step-1 is to classify the cell tower ID data into different locations to obtain the movement patterns of the user. As already discussed, the cell tower ID data that is used in this study is already classified into four different locations; i.e., H, W, E, N. This data is divided into twenty four time slots. Each time slot is represented by the associated presence information of the user (H, W, E, and N) during the one hour period as shown in Figure 4. The presence of user at specific location depends on the hour of the day and day of the week. For example, if the user has a regular routine of going to the office, then location of the user at 10a.m on Saturday morning cannot be the same at 10a.m on Monday morning. The daily life activities of an individual depend on the entropy level of the user as discussed in [3]. If the user is a low entropy user, his routines do not change much as compared to high entropy users, whose routines and activity patterns change continuously.

Step-2 is to obtain a probability matrix, which is generated depending on the hour of the day and day of the week from the classified information obtained in step-1 and then this probability matrix is used for the preparation of the training data in step-3.

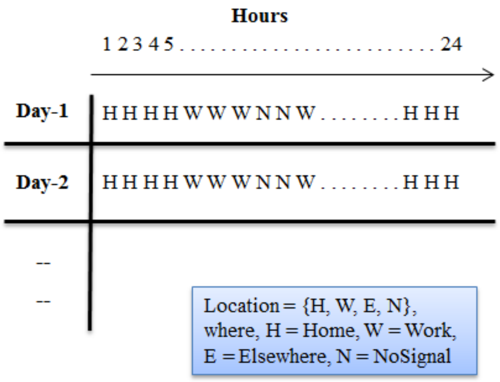


Figure 4. Format of cell tower ID Data

In step-4 decision engine detects the behaviour of the user. Two machine learning algorithms (NN and DT) are used for behaviour detection at this stage. The accuracy in terms of number of detections of both algorithms will be calculated and compared. The one, with the highest percentage accuracy will be used in the Section-6 with Bluetooth proximity data. Next section explains the behaviour analysis from cell tower ID data using NN.

A. Behaviour Analysis from Cell Tower ID Data Using Neural Networks

Figure-5 shows the basic architecture used to get the behaviour of an individual using NN. The neural network used here is Multi-layered Perceptron (MLP). Multi-layered Perceptrons have been created to try to solve the problem of non-linear classification of input instances by Rumelhart et al. [21]. A multi-layer neural network system consists of a large number of neurons connected with each other in a specific pattern. These neurons are normally divided into three classes; input layer neurons, hidden layer neurons and output layer neurons. The MLP used in this research work has four inputs and one output. Inputs are {Location, Hour, Day and Behavioural_Level}, where 'Location' gives the location of the user i.e., H, W, E, N, 'Hour' gives the hour of the day i.e., between 1 and 24, 'Day' gives the day of the week, i.e., between 1 and 7 and 'Behavioural_Level' gives the behavioural levels. Output of this neural network will give the level of abnormality of an individual for each hour of the day.

This gives twenty four samples of training data for one day. For each user, total training samples are (24 x numbers of days). 70% of these training data/samples are used for training the neural network whilst the remaining is used for cross validation and testing purposes. Training of the neural network is done till the cross validation error becomes less than 0.02, by using Mini-Batch training process [22]. The advantage of using Mini-Batch training is that it is a compromise between batch and incremental training. Back Propagation (BP) algorithm is used to estimate the weights of the neural network that includes the following steps:

- Provide a sample of training data to the NN.
- Calculate the error by comparing the desired output with the NN output.
- Adjust the weights of each neuron in order to lower the error value and again calculate the error.
- Repeat the steps unless reach the desired level.

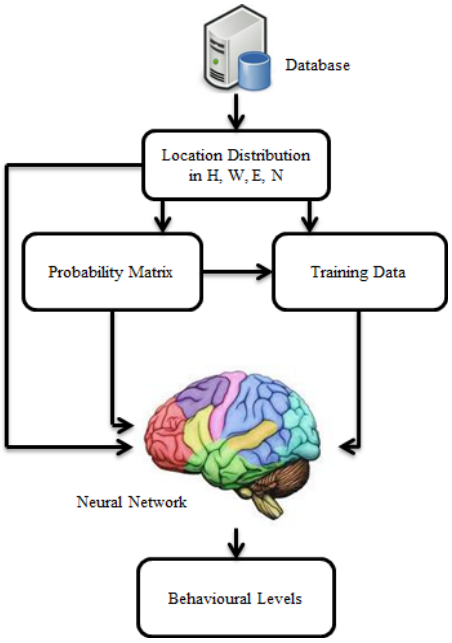


Figure 5. Behaviour Analysis from Cell Tower ID Data Using NN

Out of nine months of data available for this specific user; about 70% is used for training the neural network, one month data is used for behaviour detection and remaining is used for the cross validation purposes. Figure-6 shows the daily distributions of (H, W, E and N) transitions based on cell tower ID data of one month that is further used as a ground truth to detect the behaviour of the user and to calculate the accuracy of the NN in this specific scenario.

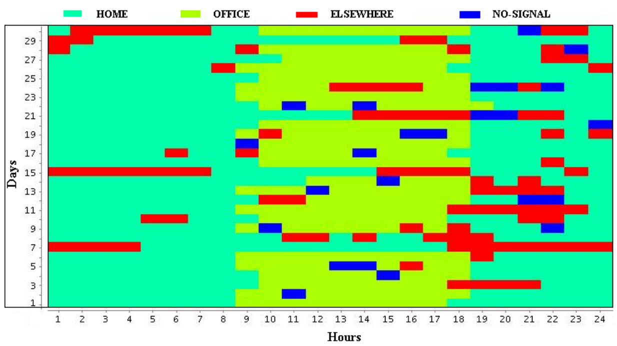


Figure 6. Distribution of (H, W, E and N) Transitions of Cell Tower ID Data

For understanding purposes, behaviour detection of the user for only two different days is discussed first. Figure-7 shows the comparison of behaviour of the user for Day-1 and Day-10. The trained neural network provides the behavioural levels for twenty four hours. First part of the figure shows the distribution of twenty four hours for Day-1 and Day-10 for the specific user in the form of H, W, E and N, whereas second part shows the inferred behaviour of the user. As the entropy level of the user is quite low, this figure shows that most of the time the behaviour of the user is average normal. Now if we look at day-10 in Figure-4, there is an unusual detection of 'Elsewhere' during 5-6am in the morning, which doesn't happen normally in usual daily routine of the user. Figure-6 also shows the detection of that unusual behaviour for day-10 in that specific time duration.

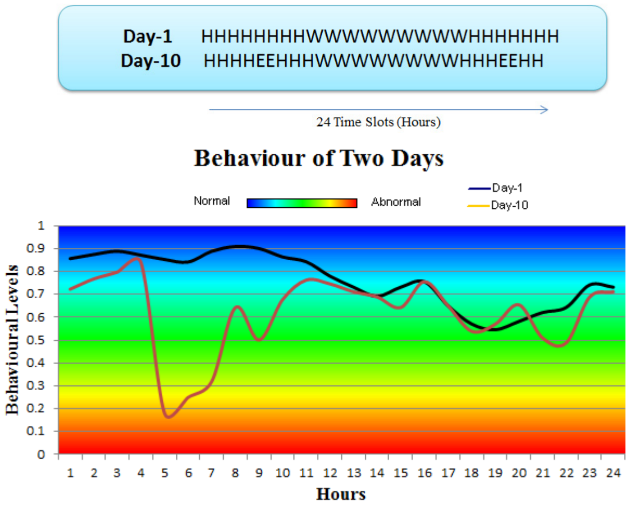


Figure 7. Comparison of Two Days of Behaviour Detected from Cell Tower ID Data Using NN

Figure-8 and Figure-9 show the behaviour of the user for one month time duration. Behaviour is divided into four levels as mentioned in Table-1. According to these levels, if the predicted behavioural value is near the '0', it means the users routine is more deviated from the normal routine activities and if it is near the '1', it is more normal. Figure-8 shows the first fifteen days and the Figure-9 shows the last fifteen days of the month. In Figure-8, the behaviour of the user for first nine days remains average normal as most of the predicted behavioural value lies between behavioural range of '0.5 - 0.7'. This can be verified from Figure-6 as well that shows the regularity in the distributions of 'Home' and 'Work' patterns and shows that user did not make any unusual movements. However, on 10th and 12th day of the month, between 5a.m – 7a.m and 10a.m – 12p.m respectively there is a change in behaviour when the user's (H, W, E and N) distributions in Figure-6 show an irregular routine activity. NN detects this behavioural change and is shown in the Figure-8 with two sharp low peaks on 10th and 12th day of the month.

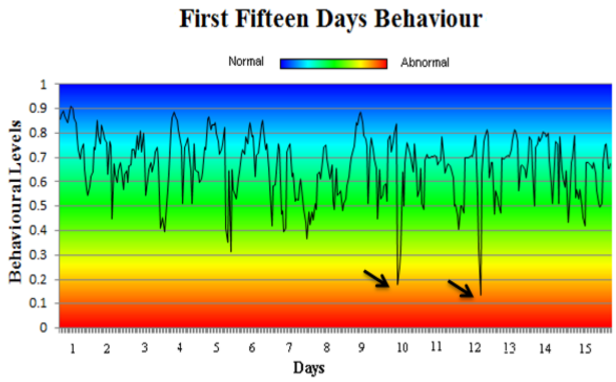


Figure 8. First Fifteen Days Behaviour Detected from Cell Tower ID Data Using NN

In Figure-9, last fifteen days of the month also show some routines that deviate from the normal behavior of the user. These routines are shown by sharp low peaks on 17th, 24th, 27th and 30th day of the month. These unusual routines are mostly detected on the week days in the morning before the office hours and some times during the office hours. As the user in these experiments belongs to academia, these results may show that, he or she most likely attending some seminar or a social function that is not part of the normal routine or due to health or traffic reasons, user sometimes comes late in the campus.

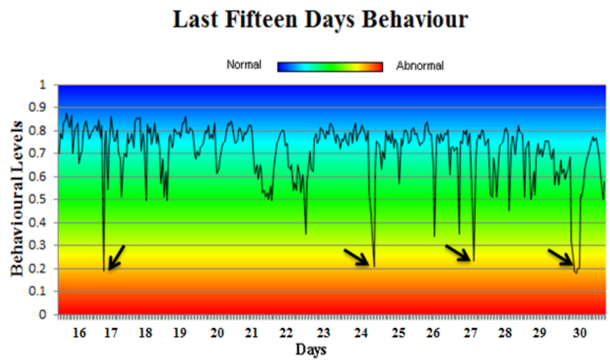


Figure 9. Last Fifteen Days Behaviour Detected from Cell Tower ID Data Using NN

B. *Behaviour Analysis from Cell Tower ID Data Using Decision Trees*

Figure-10 shows the basic architecture used to get the behaviour of an individual using DT algorithm. According to [20], DT classifies the instances by sorting them based on their feature values. Features are represented by different nodes in the DT's, and the value of the nodes is represented by the branches. Starting at the root node, each instance is classified and sorted depending upon the feature values. Root node is the feature value that best separates the data. The most well-known algorithm to build a DT is the C4.5 [23] and is used in this research work. The training and test data used for this algorithm is only the cell tower ID data of the same user as is considered in the previous section with NN. As already mentioned, this cell tower ID data is already divided into Home (H), Work (W), Elsewhere (E) and NoSignal (N) locations.

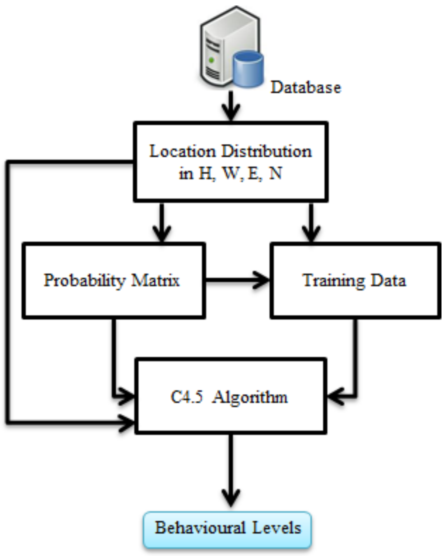


Figure 10. Data Processing Design Using C4.5 Algorithm

Figure-11 shows the behaviour detection of the user for the first fifteen days of the data using DT (C4.5 algorithm). These results show that the detections made by the DT are uniform as compared to NN, which are irregular. A DT consists of nodes and branches. Depending upon the four different behavioural levels, each node on DT represents a single behavioural level unlike NN that gives a predicted value that can be in between two different levels. The unusual routine activities are represented by the sharp low peaks in Figure-9. Another observation made is that all the unusual routines detected by DT lie in the range of 'Low_Abnormal' behavioural level and none of these are in 'Abnormal' level, unlike NN. A reason can be the biasing nature of the DT.

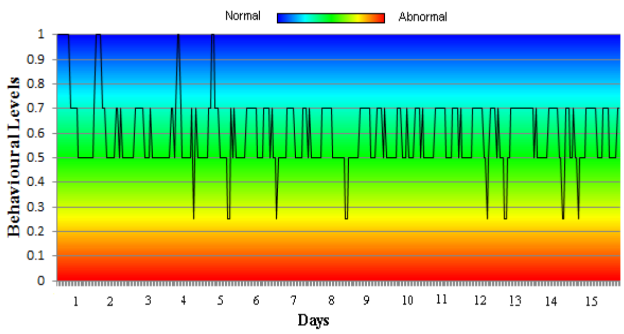


Figure 11. First Fifteen Days Behaviour Detected from Cell Tower ID Data using DT

C. Accuracy Comparison between NN and DT in detecting behaviour from cell tower

The comparison criteria for both NN and DT used here are the time required for learning and the percentage accuracy $P = N_c / N_t$, where 'N_c' is total number of correct detections and 'N_t' is the total number of detections. The results show that DT has some advantages over NN. The first advantage of induction of DT is its easy use and second advantage is that it requires fewer amounts of training data to train the classifier. The training time required for DT is also less as compared to the NN. For 3024 samples of training data and on a Pentium-IV 2.4GHz – processor with 2GB RAM, it takes only a few seconds until a decision tree has been trained. Whereas on the same system, NN takes about 13 minutes for complete training. However; the most important point for the specific problem here is, how accurate is the detection of behaviour by using DT as compared to the NN. For this purpose, percentage accuracy of both NN and DT is calculated. For 720 detections, the percentage accuracy of NN is 93% whereas DT gives the percentage accuracy of 86.3%. The bench data used for the calculation of percentage accuracy is the original data set of cell tower ID that was already classified into H, W, E and N. The reason behind the difference in the accuracy can be the biasing limitation of decision trees. As NN gives more accurate results in terms of percentage accuracy, for further processing and behaviour detection using Bluetooth proximity data, only NN will be used.

VI. BEHAVIOUR ANALYSIS FROM BLUETOOTH PROXIMITY DATA

Bluetooth proximity data is available in the form of detected devices as a result of a scanning performed by the user's cell phone after every five minutes. Each scanning results a list of devices present within the range of 5-10m. The first aim is to classify this list of detected proximate Bluetooth devices into different locations, i.e. 'Home', 'Office', 'Other Devices' and 'No Devices Found'. List of Bluetooth proximate devices does not give any direct information about the location of the user. The reason for classification of Bluetooth devices is to obtain the user's movement patterns on daily and hourly basis. By doing so, the Bluetooth data format will become same as of cell tower ID data as shown in Figure-4 and the methodology applied on cell tower ID data can be used with the Bluetooth data as well.

Another reason is that the results obtained from cell tower ID data and the Bluetooth proximity data can be compared at the end to see if we could get some interesting anomalies in behaviour of the user.

After analysing the Bluetooth proximity data, user's home computer device was given the name 'Home' (H). That means all those time slots in which user detect his home computer device, considered as 'H' because it shows user's presence in the home. For office, there are many devices that user detects during office hours. To obtain a group of devices that belong to the office, we remove the weekends from one month data and use Jaccard index [24], to detect how similar the detected devices are throughout the office hours for all remaining weekdays. Jaccard similarity equation is:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

where, 'A' and 'B' are sets of detected devices in two consecutive days. At the start, 'A' and 'B' represent day-1 and day-2, then day-2 and day-3 and so on up to the all remaining weekdays that left after removing the weekends from one month of Bluetooth data.

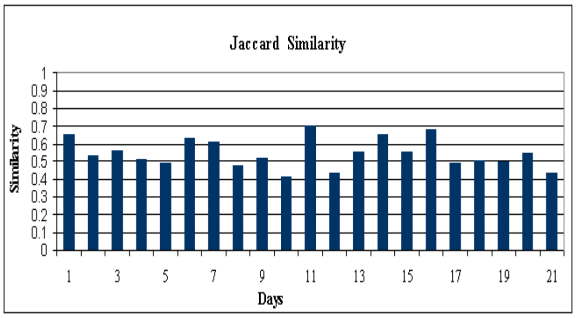


Figure 12. Jaccard Vertex Similarity

Figure-12 gives the similarity of detected Bluetooth proximate devices during the office hours between the pairs of consecutive days. The average similarity between the detected devices is above 0.5. This means there are many devices that user detects repeatedly during his office hours. All those devices that user detects for at least 70% of the days during office hours goes in ‘Office’ group. All other devices go in ‘Other Devices’ class.

After classifying the devices, a new data matrix is generated that contains twenty four time slots for each day as were in the case of cell tower ID data. Each time slot is assigned one of these classes (i.e., Home, Office, Other Devices, No Devices Found) depending upon the number of detections of the devices belonging to a specific class. Behaviour analysis frame work discussed in Section-4 is used to analyse the behaviour of low entropy user with Bluetooth proximity data using NN.

Figure-13 shows the Home/Work distribution of locations depending on the presence of user at different locations obtained from the Bluetooth data classes. The whole day is divided into twenty four time slots and each slot only represents one of the four classes depending upon the devices with which user spent most of his time.

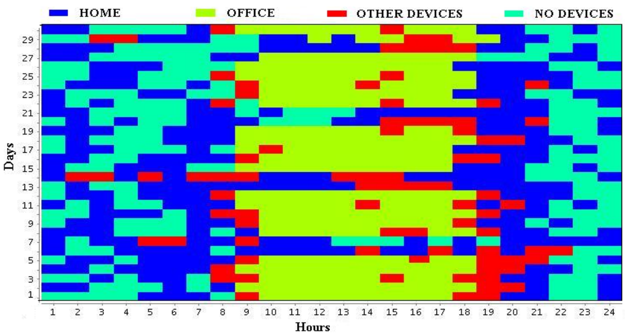


Figure 13. Distribution of Home/Work Transitions of Bluetooth Data

Figure-14 shows the fifteen days behaviour of the user detected from Bluetooth proximity data. An interesting observation can be made by analysing the behaviour detection results of both cell tower ID data in Figure-8 and Bluetooth ID data in Figure-14. It is observed that sometimes when behaviour detected from cell tower ID data is normal and no unusual pattern is detected, a change in behaviour or an unusual routine is detected from Bluetooth proximity data. For example on day-3, cell tower ID data shows normal behaviour in Figure-8, whereas an unusual routine is detected from Bluetooth proximity data on the same day shown in Figure-14. It can be said that it is more likely to be detecting unusual behaviour because during a regular routine of office hours of a weekday, user is supposed to detect ‘Office Devices’. Cell tower ID data shows this as a normal behaviour because the user is in ‘Office’ where he should be normally. Whereas Bluetooth proximity data can be pointing towards some gathering or meeting of students or staff that is not part of the regular routine. Behaviour detected from Bluetooth proximity data can be pointing towards that activity.

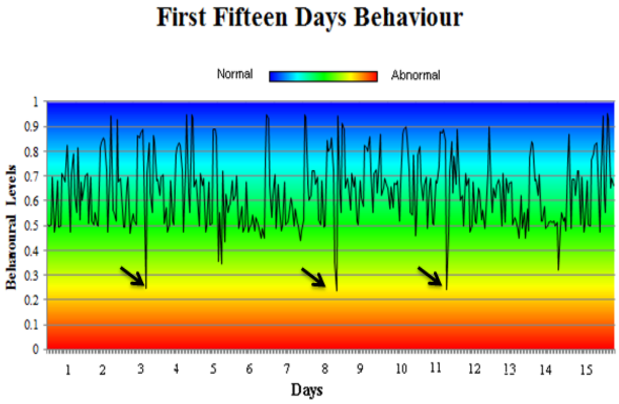


Figure 14. Fifteen Days Behaviour Using Bluetooth Proximity Data

VII. Further Analysis

So far, behaviour analysis obtained from contextual information of both GSM cell tower and Bluetooth proximity data have been presented. It is observed that the cell tower ID data gives high level behaviour of low entropy mobile people depending upon user’s location and movement in different GSM cells. Unusual behaviour in these movement patterns can be obtained from the cell tower ID data. As GSM cells cover large physical area, it can only give user’s location and does not give information about the low level activities such as attending the lecture, sitting in office with colleagues, going for shopping. On the other hand, Bluetooth proximity data gives information about other people and Bluetooth devices that are present in the close proximity of the user. It also gives information about the social relationships and most likely low level activities depending upon the detection of other proximate devices.

As the nature of contextual information obtained from cell tower and Bluetooth proximity data is different, it is interesting to analyse the difference of behaviour detected through this data statistically. For this purpose, we have used Kullback-Leibler (KL) Divergence [25], Kernel Density Estimation Function and Empirical Cumulative Distribution Function (ECDF). KL Divergence has been calculated and it gives the value of 0.4568. KL is a non-symmetric measure of the difference between two probability distributions as shown in Equation 3.

$$\mathrm{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \qquad (3)$$

where $P(i)$ and $Q(i)$ is the value obtained by taking the histogram of cell tower ID data and Bluetooth proximity data respectively. The value obtained from KL shows that there is difference in the behaviour detected by cell tower ID data and Bluetooth proximity data. This supports the argument that it is possible to use only Bluetooth proximity data to detect some behaviour of low entropy people that deviates from the normal routine, although Bluetooth doesn't give strong information about the location of the user.

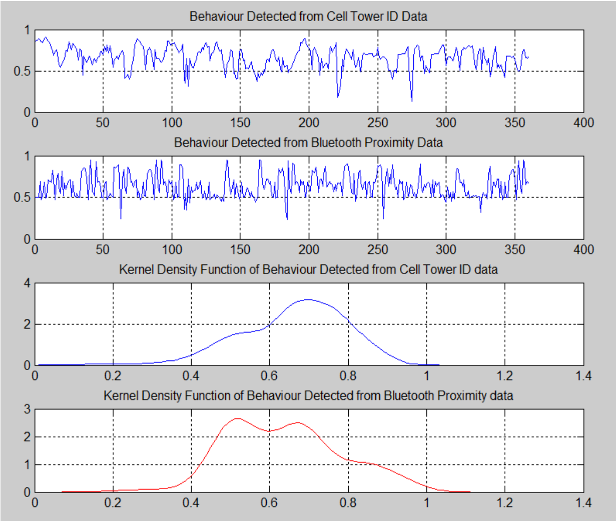


Figure 15. Kernel Density Estimation of Behaviour Detected from Cell Tower ID and Bluetooth Proximity Data

Figure-15 shows the Kernel Density Estimation [26] of two weeks of behaviour detected by both cell tower ID data and Bluetooth proximity data. Kernel Density Estimation is a non-parametric way of estimating the probability density function of a random variable. In our case, it estimates the probability density function of the behavioural values obtained by using both cell tower ID data and Bluetooth proximity data. The results in Figure-15 show that the Bluetooth proximity data show more unusual activities and routines. The peak shown around the behavioural value '0.5' in the Kernel Destiny Function of the Bluetooth data means that there are most likely many patterns in which users behaviour seems to be 'Low_Abnormal' means a little deviated from the normal routine. This shows that different routines and behaviour that deviate from the normal daily life routines of a low entropy user can be detected by using the Bluetooth proximity data.

In order to analyse difference of behaviour detected by using cell tower ID and Bluetooth proximity data in more detail, ECDF is also applied on the behavioural data. Empirical CDF is the cumulative distribution function associated with the empirical measure of the sample. Figure-16 shows the empirical distribution function applied on the behavioural data obtained using cell tower ID and Bluetooth proximity data. X-Axis shows the behavioural levels and Y-Axis gives the probability of exceeding the corresponding value on X-axis (Behavioural Levels). It shows the difference between the behaviour detected by cell tower ID and Bluetooth proximity data.

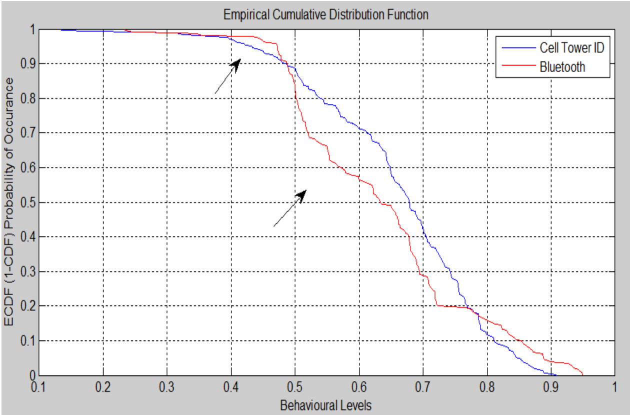


Figure 16. Empirical Cumulative Distribution Funtionc of Behaviour Detected from cell Tower ID and Bluetooth Proximity Data

Above mentioned results show that data collected from mobile devices such as cell tower ID and Bluetooth proximity data can be used for behaviour and routine activities detection. Bluetooth proximity data itself does not give much information about the location of the user but if this data is classified into different locations and user's movement patterns are obtained, then this data can give more insight into the behaviour of low entropy people. Proximity data gives information about the social relationship and activities that require social interaction of the users. If this data is used with the cell tower ID data, it can give extra information about the routines that deviates from the normal routine patterns. These results show that for low entropy users, the detection of unusual routines and behaviours by using only Bluetooth data is also possible. Low entropy users follow specific routines as compared to high entropy individuals, who live more diverse lives; therefor chances of detection of regular routines of low entropy users become more. This study aims to aid elderly people and patients to detect abnormal and unusual behaviours to avoid any accidents. Normally patients and elderly people have fixed and limited routines to follow that can likely be detected using Bluetooth devices by classifying the Bluetooth Proximity data into different activities or communities.

VIII. Summary and Future Work

In this paper, real time Bluetooth proximity and cell tower ID data is used to detect activities and routines of an individual that deviates from the normal daily life routines by using NN and DT. A low entropy user was selected for experiments due to the regularity and constancy in his routines. A successful detection of abnormal behaviour in this user's routines is done by using cell tower ID's and Bluetooth proximity data. NN and DT are used as decision engines to detect the behaviour of the user by using cell tower ID data. NN are found more accurate as compared to the DT in detection. However; DT requires less data and takes less time for training.

Bluetooth proximity data is classified into four different categories by using Jaccard Index. To detect anomalies in more specific and lower level activities and routines, we need to classify the Bluetooth proximity data into temporal clusters. In future work, we will try to classify the Bluetooth proximity data on temporal scale to cover the minute details of the user's behaviour and will also try to predict the behaviour based on these classes and communities detection using pervasive computing`. This will provide one step further in the identification of unusual routines and activities by using only Bluetooth proximity data. This will help us to facilitate elderly people and patients who need more care and concern about their behaviour and unusual routines that can cause serious accidents.

ACKNOWLEDGMENT

This research work is funded by the Higher Education Commission of Pakistan under Faculty Development Program through University of Engineering & Technology Taxila, Pakistan.

REFERENCES

[1] Azam, M. A. Tokarchuk, L. Adeel, M. *"Human Behaviour detection Using GSM Location Patterns and Bluetooth Proximity Data"*. The Fourth International Conference on Mobile Ubiquitous Computing, Services and Technologies, pp. 428-433, Florence, Italy, 2010.

[2] Hermersdorf, M. Nyholm, H. Perkio, J. Tuulos, V. *"Sensing in Rich Bluetooth Environments"*- Workshop on WorldSensorWeb, in Proc. SenSys, 2006 - sensorplanet.org

[3] Eagle, N. Pentland, A. *"Reality mining: sensing complex social systems*". Personal and Ubiquitous Computing 2006 – Springer, Vol. 10, # 4, 255-268

[4] Gill, T. M. Desal, M.M. Evelyne, A. Holford, T. R. Williams, C. S. *"Restricted-Activity among Community-Living Older Persons: Incidence, Participants, and Health Care Utilization"*, Annals of Internal Medicine. http://www.annals.org/content/135/5/313.full.pdf+html

[5] Azam, M. A. Laurissa, T. (2009). *"Behaviour Detection Using Bluetooth Proximity Data";*. Proceedings of Networking & Electronic Commerce Research Conference, pp. 46-52 (NAEC 2009).

[6] Hara, K. Omori, T. Ueno, R. *"Detection of unusual human behaviour in intelligent house"*; Proceedings of the 2002 12th IEEE workshop on Neural Networks for Signal Processing, pp. 697-706, 2002.

[7] Yiping, T. Zhiying, Z. Hui, G. Huiqiang, L. Wei, W. Gang, X. *"Elder Abnormal Activity Detection by Data Mining"*, SICE Annual Conference in Sapporo, August 4-6, 2004, vol. 1, pp. 837–840 (2004) Japan

[8] Wren, C. Ivanov, Y. Kaur, I. Leigh, D. Westhues, J. "*SocialMotion: Measuring the Hidden Social Life of a Building"*. In: J. Hightower, B. Schiele, and T. Strang, (eds.) LoCA 2007. LNCS, vol. 4718, pp. 85–102. Springer, Heidelberg (2007)

[9] McCowan, I. Gatica-Perez, D. Bengio, S. Lathoud, G. *"Automatic Analysis of Multimodal Group Actions in Meetings"*. IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI) 27(3), 305–317 (2005)

[10] Stiefelhagen, R. Bernardin, K. Ekenel, H.K. McDonough, J. Nickel, K. Voit, M. Woelfel, M. *"Audio-Visual Perception of a Lecturer in a Smart Seminar Room*". In: Signal Processing - Special Issue on Multimodal Interfaces, vol. 86 (12). Elsevier, Amsterdam (2006)

[11] Eagle, N. "*Machine Perception and Learning of Complex Social Systems",* Ph.D. Thesis, Program in Media Arts and Sciences, MIT, June 2005

[12] Nicolai, T. Behrens, N. Yoneki, E. *"Wireless Rope: Experiment in social proximity sensing with Bluetooth"*. In fourth annual IEEE International Conference on Pervasive Computing, LNCS 4277, 2006

[13] Farrahi, K. Gatica-Perez, D. *"Daily Routine Classification from Mobile Phone Data*". In: Popescu-Belis, A., Stiefelhagen, R. (eds.) MLMI 2008. LNCS, vol. 5237, pp. 173–184. Springer, Heidelberg (2008)

[14] Farrahi, K. Gatica-Perez, D. *"What did you do today? Discovering daily routines from Large-Scale Mobile Data"*. In: MM 2008: Proceeding of the 16th ACM International Conference on Multimedia, pp. 849–852. ACM, New York (2008)

[15] Vukovic, M. Vujnovic, G. Grubisic, D. *"Adaptive User Movement Prediction for Advanced Location-aware Services"*, Proceedings of the 17th international conference on Software, Telecommunications and Computer Networks, pp. 343-347 (2009)

[16] Vukovic, M. Lovrek, I. Jevtic, D. "*Predicting user movement for advanced location-aware services"*. In 15th International Conference on Software, Telecommunications and Computer Networks, pp. 1–5. SoftCOM 2007, 2007.

[17] Tabia, K. Benferhat, S. *"On the Use of Decision Trees as behavioural Approaches in Intrusion Detection"*. In Seventh International Conference on Machine Learning and Applications, 2008. ICMLA'08.

[18] Hu, D. H. Zhang, X. Yin, J. Zheng, V.W. Yang, Q. *"Abnormal Activity Recognition Based on HDP-HMM Models",* AAAI Publications, 21st International Conference on Artificial Intelligence, pp. 1715–1720, 2009

[19] Eagle, N. Clauset, A. Quinn, J. A. *"Location Segmentaion, Inference and Prediction for Anticipatory Computing",* AAAI Spring Symposium, 2009

[20] Kotsiantis, S. *"Supervised Machine Learning: A Review of Classification Techniques*", Informatica Journal 31, pp. 249-268, 2007.

[21] Rumelhart, D. E. Hinton, G. E. Williams, R. J. *"Learning Internal Representations by Error Propapgation"*. In: Rumelhart D. E., McClelland J L et. al. (eds) Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT press, Cambridge, MA, Vol. 1, pp. 318-362.

[22] Linear Neural Networks, retrieved on 03-09-2011, retrieved from http://www.idsia.ch/NNcourse/linear2.html

[23] Quinlan, J. R. "*C4.5: Programs for Machine Learning*". Morgan Kaufmann San Francisco.

[24] Jaccard Index, retrieved on 03-09-2011, retrieved from http://www.statemaster.com/encyclopedia/Jaccard-index

[25] Kullback, S. Leibler, R. A. (1951). "*On Information and Sufficiency*". Annals of Methematical Statistics 22(1): 79-86.

[26] Rosenblatt, M. (1956). "*Remarks on some nonparametric estimates of a density function*". Annals of Mathematical Statistics 27: 832–837.