On the Prediction of Future User Connections Based on Historical Records in Wireless Networks

Seyedeh Soheila Shaabanzadeh, Juan Sánchez-González

Universitat Politècnica de Catalunya (UPC), c/ Jordi Girona, 1-3, 08034 Barcelona, Spain
`juansanchez@tsc.upc.edu`

Abstract. Recent developments of data monitoring and analytics technologies in the context of wireless networks will boost the capacity to extract knowledge about the network and the users. On the one hand, the obtained knowledge can be useful for running more efficient network management tasks related to network reconfiguration and optimization. On the other hand, the extraction of knowledge related to user needs, user mobility patterns and user habits and interests can also be useful to provide a more personalized service to the clients. Focusing on user mobility, this paper presents a methodology that predicts the future Access Point (AP) that the user will be connected to in a Wi-Fi Network. The prediction is based on the historical data related to the previous APs which the user connected to. Different approaches are proposed, according to the data that is used for prediction, in order to capture weekly, daily and hourly user activity-based behaviours. Two prediction algorithms are compared, based on Neural Networks (NN) and Random Forest (RF). The methodology has been evaluated in a large Wi-Fi network deployed in a University Campus.

Keywords: Wi-Fi Network, Supervised Learning, Length of Stay Prediction, Neural Network (NN), Random Forest (RF).

1 Introduction

The demand of new multimedia services (i.e. online multimedia applications, high quality video, augmented/virtual reality, etc.) has dramatically increased in the last years. A solution to cope with the high bandwidth and strict Quality of Service (QoS) requirements associated to these new services consists on network densification through the deployment of Small Cells (SC) operating cellular technologies (e.g. 4G/5G), complemented with Wi-Fi hotspots using the unlicensed spectrum. In fact, Wi-Fi technology is a competitive option for serving multimedia demands due to its popularity among mobile users. In the last years, a dramatic increase in the amount of IEEE 802.11 (i.e. Wi-Fi) traffic has been observed. It is estimated that by 2021, 63% of the global cellular data traffic will be offloaded to Wi-Fi or small cell networks [1]. Globally, there will be nearly 549 million public Wi-Fi hotspots by 2022, up from 124 million in 2017, a fourfold increase.

In the last years, we have witnessed a widespread use of powerful (Big) Data monitoring and analytics technologies in many areas of our lives. In particular, in the context of cellular and Wi-Fi networks, the use of these technologies will be one of

the main pillars to cope with the above-mentioned challenges. The monitoring system provides the ability to collect information about the users and the network, while the analytics system allows to extract knowledge of the collected data by means of Artificial Intelligence (AI) mechanisms [2]. There are different ways to extract knowledge (e.g. by using classification, clustering and prediction mechanisms) [3].

Multiple applicability examples of data monitoring and analytics in the context of Wi-Fi or cellular networks can be found in the literature. As an example related to mobile cellular networks and concerning the characterization of user habits, the collection of the Base Station and the mobile terminal communication activity (messages, calls, etc.) has been used for urban and transportation planning purposes to identify daily motifs, given that human daily mobility can be highly structural and organized by a few activities essential to life [4]. Similarly, [5] proposed a methodology to partition a population of users tracked by their mobile phones into four predefined user profiles: residents, commuters, in transit and visitors. Applications envisaged are traffic management, to better understand how traffic is affected by the residents mobility compared to the commuters, or studying how the city is receiving people from outside and how their movements affect the city. [6] proposed an agglomerative clustering to identify user's daily motifs according to the cells in which the user is camping during the day. Real measurements obtained from a 3G/4G network were used.

The obtained knowledge related to the network status, the performance of the services, user habits, user requirements, etc. can also be useful for supporting different decision-making processes over the network (e.g. adjusting the usage of the network resources) which will lead to more efficient network management tasks related to network reconfiguration and optimization. As an example, the use of prediction methodologies for identifying the future SC/AP which the user will be connected to, together with an estimation of the future user traffic volume and perceived user performance may provide a more accurate future user characterization. This can be useful for carrying out a more proactive network reconfiguration approach. In the context of mobility management in 5G cellular networks, anticipating the cell to which users will be connected in the future is useful to facilitate handover procedures [7]. In the context of Wi-Fi networks, the prediction of APs to which a client will connect in the future can be useful for a more efficient Pairwise Master Key (PMK) caching or Opportunistic Key Caching (OKC) techniques which can reduce the time for re-authentication when roaming to a new AP [8]. Concerning the AP selection in Wi-Fi networks, [9] proposed a mobility prediction methodology based on a Hidden Markov Model that is used to forecast the next AP that users will connect to, based on current and historical user location information. On the other hand, from the point of view of traffic offloading from cellular to Wi-Fi systems, length of stay prediction at an AP can be useful for user bandwidth allocation e.g. giving higher priority to soon-to-depart Wi-Fi users so that the larger amount of traffic is sent through the Wi-Fi before performing the handover to the cellular system [10].

The extracted knowledge related to user location and mobility, user habits and interests, etc. can also be useful for companies for commercial purposes. As an example, a reliable prediction of future user location and length of stay connected to the

different SC/AP enables the use of Location Based Advertising (LBA) mechanisms. This presents the possibility for advertisers to personalize their messages to people based on their location and interests [11].

Within this context, this paper proposes a methodology that predicts the future user connections to the different APs of a Wi-Fi network according to historical user records. The main contribution of the paper is the proposal of a prediction methodology that is able to extract user periodical patterns at different time-levels in order to capture weekly, daily and hourly user activity-based behaviours. Two prediction algorithms based on a supervised learning process are compared, one using a Neural Network (NN) and the other one based on Random Forest (RF). The proposed methodologies are evaluated for a large Wi-Fi network deployed in a University Campus. The remaining of the paper is organized as follows. Section 2 presents the proposed AP prediction methodology, while Section 3 describes the considered prediction tools. The results are presented in Section 4, while Section 5 summarizes the conclusions.

2   Proposed Methodology

The proposed prediction methodology is shown in Fig. 1 and assumes a Wi-Fi Network with monitoring capabilities for the collection of measurements reported by the users when connected to the different APs. The *Collection of Network Measurements* process collects a list of metrics for each u-th user ($u=1,…,U$) when connected to each AP (e.g. the instants of time when the user begins and ends a connection to each AP, the average SNR -Signal to Noise Ratio-, the average RSSI -Received Signal Strength Indicator-, the amount of bytes transmitted/received during the connection of the user to each AP, etc.). All this information is stored in a database. Then, for each user, a pre-processing of the collected data is done so that the measurements collected during each d-th day (with $d=1,…,D$) are grouped in M time periods with equal duration T. In particular, the *pre-processing* step generates a matrix A for each user, so that each term $a_{d,m}$ (with $m=1,…,M$ and $d=1,…D$) represents the AP identifier to which this user was connected during the m-th time period of the d-th day. In case that the user connects to more than one AP at the same m-th time period, it is assumed that the term $a_{d,m}$ will correspond to the AP with the highest connection duration.

For the prediction of the AP to which the user will be connected in a specific m^*-th time period of a specific d^*-th day in the future ($a_{d*,m*}$), the proposed methodology makes use of some historical information of the AP to which the user was connected in the past and a prediction function $f(\cdot)$ that is obtained by means of a supervised learning. For that purpose, the *Selection of historical data* process selects some specific terms in matrix A. Different approaches are presented below:

- *Prediction Based on Time-period Patterns (PBTP):* In this case, the prediction of $a_{d*,m*}$ is based on the APs to which the user was connected in the last N previous time periods (i.e. $a_{d*,m*-N},…,a_{d*,m*-n},…,a_{d*,m*-1}$). In order to obtain the prediction function $f(\cdot)$, a vector $B=(b_1,…,b_f,…b_F)$ is built from all the $F=D·M$ previous time periods in the last D days, i.e. $B=(a_{1,1},…,a_{1,m},…,a_{1,M},…,a_{d,1},…,a_{d,m},…,a_{d,M},…,a_{D,1},…,a_{D,m},…,a_{D,M})$.

Then, B is split into I different training tuples B^i, each one composed of the i-th element and its N previous elements. This split is done by applying a sliding window of length N over the set of F measurements, resulting in a training set of F-N training tuples of the form $B^i=(b_{i-N},...,b_{i-n},...,b_i)$ with $i=1,...,F$-N. The $f(\cdot)$ function is learnt by a supervised learning that consists on observing the relationship between b_i and its N previous elements $(b_{i-N},...,b_{i-1})$ for all the $I=F$-N tuples. The rationale of this approach is to identify user frequent AP connectivity patterns in N consecutive time periods.

   - *Prediction Based on Daily Patterns (PBDP)*: In this case, the prediction of $a_{d*,m*}$ is done according to the APs to which the user was connected in the previous N days at the same time period of the day (i.e. $a_{d*-N,m*},...,\ a_{d*-n*,m},...,\ a_{d*-1,m*}$). In order to obtain the prediction function $f(\cdot)$, a set of M vectors $B_m=(b_{1,m},...,b_{f,m},...,b_{F,m})$ is built. Each B_m consists on the AP to which the user connected in the last $F=D$ days at the m-th time period of the day, i.e. $B_m=(a_{1,m},...,a_{D-d,m},...,a_{D-1,m},a_{D,m})$. Then, each B_m is split into I different training tuples B^i_m, each one composed of the i-th element and its N previous elements, i.e. $B^i_m=(b_{i-N,m},...,b_{i-n,m},...,b_{i,m})$ with $i=1,...,D$-N. Then, a total of $(D$-$N)\cdot M$ tuples with size $N+1$ are generated. The rationale of this is to identify user periodical AP connectivity patterns in N consecutive days at the same time of the day.

   - *Prediction Based on Weekly Patterns (PBWP):* In this case, the prediction of $a_{d*,m*}$ is done according to the APs to which the user was connected in the N previous weeks at the same day of the week and time period of the day (i.e. $a_{d*-7N,m*},...,$ $a_{d*-7n,m*},...,a_{d*-7,m*}$). Again, a set of M vectors $B_m=(b_{1,m},...,b_{f,m},...,b_{F,m})$ is built, each one consisting on the AP to which the user connected in the last $F=W$ weeks at each m-th time period of a specific day of the week, i.e. $B_m=(a_{d-7W,m},...,\ a_{d-7w,m},...,a_{d-7,m},a_{d,m})$. Then, each B_m is split into different I training tuples B^i_m, each one composed of the i-th element and its N previous elements, i.e. $B^i_k=(b_{i-N,m},...,b_{i-n,m},...,b_{i,m})$ with $i=1,...,W$-N. In this case, the total number of tuples is $(W$-$N)\cdot M$. The rationale of this is to identify weekly AP connectivity patterns at the same day of the week and time of the day.

   - *Joint Based Prediction (JBP)*: It consists on the same methodology but doing a combination of the three approaches described previously. In this case, the prediction of $a_{d*,m*}$ is done according to the APs to which the user connected in the last N time periods (i.e. $a_{d*,m*-N},...,a_{d*,m*-n},...,a_{d*,m*-1}$), the APs at the same m-th time period for the last N days (i.e. $a_{d*-N,m*},...,a_{d*-n,m*},...,a_{d*-1,m*}$) and the APs at the same day of the week and time period of the day for the last N weeks (i.e. $a_{d*-7N,m*},...,a_{d*-7n,m*},...,a_{d*-7,m*}$). The prediction function $f(\cdot)$ is learnt in a similar way as before by observing the relationship of specific $a_{d,m}$ and the observations in the last N time periods, the last N days at the same time period and the last N weeks at the same day of the week and time period of the day. A total number of $(W$-$N)\cdot M$ tuples with size $3N+1$ are obtained.
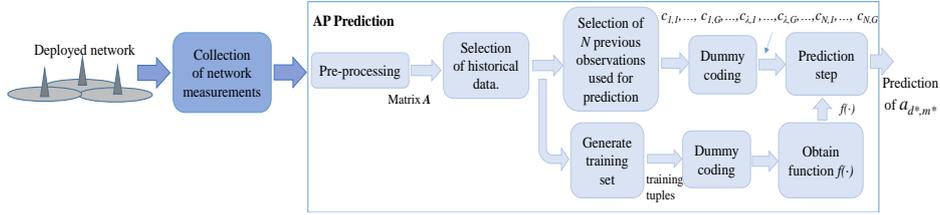
Fig. 1. Proposed prediction methodology.

It is worth noting that all these terms $a_{d,m}$ in matrix $\boldsymbol{A}$ correspond to categorical values (e.g. an AP identifier). For both the training and prediction, these terms are converted into numerical attributes by means of the so-called dummy coding process [11]. A dummy variable is a binary variable coded as 0 or 1 to represent the absence or presence of some categorical attribute. Therefore, each of the N elements used for prediction a_λ ($\lambda=1,...,N$) are converted into a set of G dummy variables $\boldsymbol{c}_\lambda=(c_{\lambda,1},...,c_{\lambda,g},...,c_{\lambda,G})$, where G is the number of different APs in the set of N measurements, so that the term $c_{\lambda,g}=1$ if a_λ corresponds to the g-th AP and $c_{\lambda,g}=0$ otherwise. Then, the resulting number of dummy variables $D=N\cdot G$ are used for the prediction of $a_{d*,m*}$ according to (1). Before the training, the same dummy coding is also done for all the training tuples of the training set.

$$a_{d*,m*}=f(c_{1,1}, ..., c_{1,G}, ...,c_{\lambda,1}, ...,c_{\lambda,G}, ...,c_{N,1}, ..., c_{N,G}) \tag{1}$$

3   Considered Supervised Learning Algorithms

In this paper, a supervised learning algorithm based on Neural Networks and Random Forest are compared. A brief description of them is provided below.

Neural Networks (NN): In this case, the prediction is done by means of a feedforward Neural Network that consists on an input layer, one or more hidden layers and an output layer [12]. Each layer is made up of processing units called neurons. The inputs are fed simultaneously into the units of the input layer. Then, these inputs are weighted and are fed simultaneously to the first hidden layer. The outputs of the hidden layer units are input to the next hidden layer, and so on. A supervised learning technique called backpropagation is used for training. Back propagation iteratively learns the weights of the Neural Network by comparing the inputs and outputs of the training set.

Random Forest (RF): Ensemble methodologies are used to increase the overall accuracy by learning and combining a series of individual classifier models. Random Forests is a popular ensemble method. RF is based on building multiple decision trees, generated during the training phase, and merge them together in order to obtain a more accurate and stable prediction [3]. Different from the single decision tree methodology, where each node of the tree is split by searching for the most important feature, in random forest, additional random components are included (i.e. for each node of each tree the algorithm searches for the best feature among a random subset

of features). Once all the trees are built, the result of the prediction corresponds to the most occurred prediction from all the trees of the forest.

4 Results

The considered scenario consists on a large Wi-Fi network with 429 APs deployed in a University Campus with 33 buildings with four floors per building. The reported user measurements are collected by the Cisco Prime Infrastructure tool [13]. The users' measurements were collected during $D=84$ consecutive days (i.e. $W=12$ weeks). The prediction methodology was run for $U=967$ users. According to the methodology described in Section II, the matrix A is built for each user by determining the AP to which the user is connected in each of the $M=96$ periods of $T=15$ minutes for each of the $D=84$ days. According to this data, the proposed prediction methodology is run in order to predict the AP which each user will be connected to in all the $M=96$ time periods of $T=15$ minutes in the subsequent week. The obtained predictions are compared to the real APs which the user connected to. The prediction accuracy is calculated as the percentage of time periods that have been predicted correctly in the range between 6:00h and 22:00h for all the weekdays (from Monday to Friday) for all the users that connected to the Wi-Fi network at least one time every day. The prediction methodology has been implemented by means of Rapidminer Studio 14]. The parameters of each supervised learning algorithm have been tuned to obtain the maximum prediction accuracy. In particular, the Neural Network is configured with learning rate 0.05, momentum 0.9, 100 training cycles and 1 hidden layer of size 20 and the Random Forest has been configured with 100 trees, gain_ratio criterion and maximal depth 10.

4.1 Example of the AP prediction methodology

In order to illustrate the performance of the proposed prediction methodology, let first focus on the AP prediction for a specific user for all the time periods on a Wednesday. Assuming here the PBWP approach, the AP prediction at the m-th time period is based on the APs to which the user connected to in the previous $N=6$ Wednesdays at the same m-th time period of the day. The training set is built by using the last $F=12$ weeks. For validation purposes, the predictions are compared to the real AP where the user connected to during this Wednesday. According to this, Fig. 2 presents this comparison when using the Neural Network algorithm. As shown, for this particular user, the methodology is able to correctly predict the AP in 60 out of 64 periods of 15 minutes (i.e. a 93.75% of prediction accuracy). In general, most of the transitions between AP are correctly predicted. In fact, only an error of one period of 15 minutes is observed in predicting the user time of arrival while a slightly higher error is also observed for the prediction of departure.
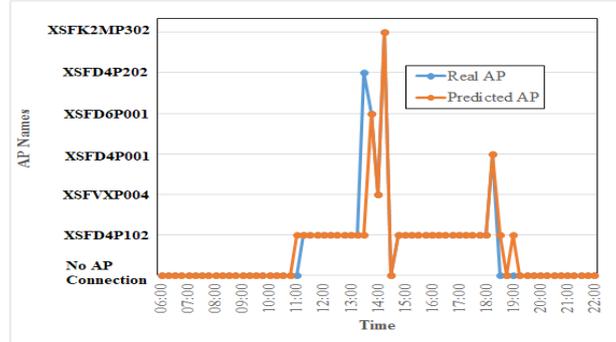
Fig. 2. Comparison of real and predicted AP for a specific user for Wednesday with NN.

During lunch time, the connection to AP XSFD4P202 is not correctly-predicted, but the prediction was AP XSFD4P102 that is located just in the lower floor of the same building. This indicates that, although the AP is not well predicted in this case, the methodology predicted correctly the region where the user was located.

4.2   Comparison of the different proposed approaches.

In order to gain insight in the performance of the proposed methodology, the prediction process has been run for all the set of users in all the time periods of 15 minutes during a whole week. Then, the predictions were compared to the AP to which each user connected at each time period. Table 1 presents the percentage of users in which each approach provides the best prediction accuracy for PBWP, PBDP and PBTP. As shown for both NN and RF predictors, PBTP approach provides better prediction accuracy than PBWP or PBDP for most of the users. This indicates that the AP to which a user was connected in the most recent time periods is the most useful information for prediction. However, it is worth noting that, for a relatively high percentage of users, the best approach is obtained with PBWP or PBDP (e.g. around 30% and 18% for NN and RF, respectively). This result indicates that the daily or weekly periodical behavior of some users can be captured better by PBDP or PBWP approaches, respectively.

Table 1. Percentage of users in which the different approaches provide the best accuracy.

	NN	RF
PBWP	24.07	18.62
PBDP	6.01	0
PBTP	69.90	81.37

Fig. 3 presents the Cumulative Distribution Function of the prediction accuracy for the different prediction approaches with Neural Network. For comparison purposes, in all the approaches, the prediction is based on the 6 previous observations. Therefore, in PBWP, PBDP and PBTP, the sliding window is set to $N=6$. In JBP approach, the sliding window is set to $N=2$, i.e. the prediction is based according to the APs to

which the user connected to in the last $N=2$ time periods (i.e. $a_{d*,m*-2}$, $a_{d*,m*-1}$), the APs at the same m-th time period for the last $N=2$ days (i.e. $a_{d*-2,m*}$, $a_{d*-1,m*}$) and the APs at the same day of the week and time period of the day for the last $N=2$ weeks ($a_{d*-14,m*}$, $a_{d*-7,m*}$). As shown in Fig. 3, the JBP approach is able to provide a better prediction accuracy than the rest of the approaches separately. The reason is that the JBP is able to jointly capture the hourly, daily and weekly user behavior.

In Table 2, the average prediction accuracy and the average computation time, required for running the methodology for each user, are compared for the different approaches for both NN and RF. The methodology was executed in a computer with a Core i5-3330 processor at 3.00 GHz and RAM memory of 8GB running Microsoft Windows 10. It has been observed that the computation time is mainly due to the process of training while the time for the prediction step is negligible. As shown in Table 2, the PBTP and PBDP approaches exhibit higher computation time per user since they make use of larger number of training tuples, leading to longer training times. As shown in Table 2, the JBP approach provides the best prediction accuracy with a relatively low computation time.
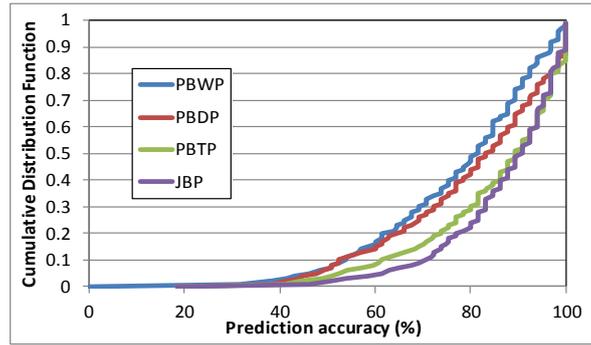

Fig. 3. CDF of the prediction accuracy the different approaches.

Table 2. Comparison of the different approaches.

	Average prediction accuracy (%)		Computation time per user (s)	
	NN	RF	NN	RF
PBWP	77.16	79.29	0.87	0.22
PBDP	79.88	79.64	33.33	3.74
PBTP	84.84	84.53	38.95	3.94
JBP	86.73	84.26	3.21	0.64

It is worth noting that the computation time required for the training may impose some restrictions in the maximum number of users that can included in the AP prediction or the frequency in which the training is updated. The values of the average computation time per user obtained in Table 2 may be excessively high in a Wi-Fi network that may have several thousands of simultaneous user connections. As a consequence, running the proposed methodology for such a high amount of users may re-

quire the parallelisation of the proposed methodology using multiple processors, each of them to run the prediction of a group of users.

4.3   Impact of the amount of data used for training

This section presents the impact of the amount of historical data used for building the training set. In particular, Fig. 4a presents a comparison of the prediction accuracy for JBP with *N=2* for Neural Network and Random Forest as a function of the number of days with measurements considered for generating the training set. On the other hand, Fig. 4b shows the average computation time required per user. As shown in Fig. 4a, the Neural Network predictor provides higher average prediction accuracy than Random Forest. Fig. 4a also shows that the use of larger amount of data for generating the training set provides higher prediction accuracy. However, processing larger amount of data requires higher computation time, especially for the Neural Network, as shown in Fig. 4b.
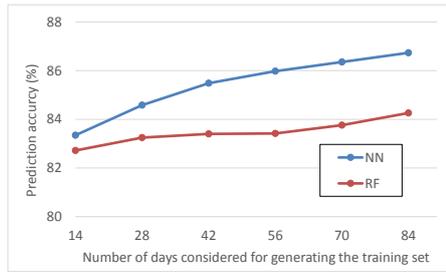

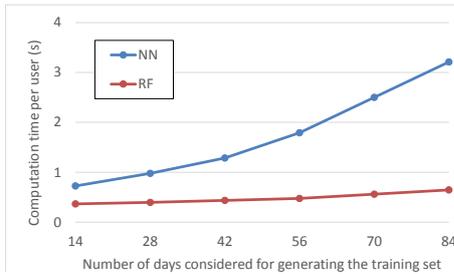**Fig. 4a.** Average prediction accuracy.   **Fig. 4b.** Average Computation Time per user.

4.4   Impact of the size of the sliding window

This section evaluates the impact of the size of the sliding window for the JBP approach when *D=84* days of measurements are considered for generating the training set. In particular, Table 3 presents the average prediction accuracy and the computation time per user for different values of the sliding window *N*. As shown in Table 3, a too low value of the sliding window reduces the capability to detect weekly, daily and hourly user periodical behaviour, which leads to a lower prediction accuracy. However, when setting a too high sliding window, the number of the tuples for generating the training set will become lower and, as a consequence, a worse training process is done, which reduces the prediction accuracy.

Table 3. Impact of the size of the sliding window

	Average prediction accuracy (%)		Computation time per user (s)	
	NN	RF	NN	RF
N=2	86.73	84.26	3.21	0.64
N=6	90.71	84.75	4.75	0.79
N=10	88.79	82.96	2.54	0.71

5 Conclusions

This paper has proposed a methodology for the prediction of future APs to which users will be connected in a Wi-Fi network. The proposed methodology is based on a supervised learning that makes use of historical user connectivity to build a prediction model. Different approaches have been defined depending the historical data that is used. In general, the PBTP approach, in which the prediction is based according to the most recent APs to which the user connected, provides the best prediction accuracies. However, PBDP or PBWP perform better for users that follow some daily or weekly periodical behavior. As shown, a joint approach (JBP) is able to provide better prediction accuracy than the rest of the approaches separately with a relatively low computation time per user. The impact of the training set size has been illustrated for the JBP approach in terms of prediction accuracy and computation time. As shown, higher amount of days with measurements considered for generating the training set provides higher prediction accuracy at expenses of higher computation time, especially for Neural Networks. The impact of the size of the sliding window has been also evaluated. A too low value of the sliding window results in a worse capability to detect weekly, daily and hourly user periodical behavior while a too large value leads to a too low number of tuples for training. The results indicate that the prediction based on the Neural Network provides a higher prediction accuracy than the prediction based on Random Forest at expenses of an increase in the computation time.

References

1. Cisco Visual Networking Index, "Forecast and Trends, 2017-2022", Cisco white papers November, 2018.
2. I. C-L, Y. Liu, S. Han, S. Wang, G. Liu, "On Big Data Analytics for Greener and Softer RAN", *IEEE Access*, August, 2015.
3. R. Nisbet, G. Miner, K. Yale, *Handbook of Statistical Analysis and Data Mining Applications*, 2nd edition, Academic Press, November 2017.
4. S. Jiang, J. Ferreira, M.C. González, "Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore", *IEEE Transactions on Big Data*, 2015.
5. B. Furletti, L. Gabrielli, S. Rinzivillo, C. Renso, "Identifying users' profiles from mobile calls habits", *International Workshop on Urban Computing*, 2012.
6. J. Sánchez-González, O. Sallent, J. Pérez-Romero, R. Agustí, "On Extracting User-centric Knowledge for Personalised Quality of Service in 5G Networks", *International Symposium on Integrated Network Management*, 2017.
7. 3GPP TR 23.791 "Technical Specification Group Services and System Aspects. Study of Enablers for Network Automation for 5G".
8. A. Kumar, H. Om, "A Secure Seamless Handover Authentication Technique for Wireless LAN", *International Conference on Information Technology*, 2015.

9. Y. Khong-Lim, C. Yung-Wey, "Optimised Access Point Selection with Mobility Predicion Using Hidden Markov Model for Wireless Networks", *International Conference on Ubiquitous and Future Networks (ICUFN)*, 2017.

10. J. Manweiler, N. Santhapuri, R.R. Choudhury, S. Nelakuditi, "Predicting Length of Stay at WiFi Hotspots", *IEEE International Conference on Computer Communications (INFOCOM)*, April 2013.

11. C. Yan, P. Wang, H. Pang, L. Sun, S. Yang, "CELoF: WiFi Dwell Time Estimation in Free Environment", *MultiMedia Modeling MMM 2017. Lecture Notes in Computer Science*, Springer, 2017.

12. J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufman (MK), Elsevier, 2006.

13. *Cisco Prime Infrastructure 3.5 Administrator Guide*, available at www.cisco.com

14. RapidMiner Studio, http://www.rapidminer.com