On Learning and Exploiting Time Domain Traffic Patterns in Cellular Radio Access Networks

Jordi Pérez-Romero, Juan Sánchez-González, Oriol Sallent, and Ramon Agustí

Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

{jorperez, juansanchez, sallent, ramon}@tsc.upc.edu

Abstract. This paper presents a vision of how the different management procedures of future Fifth Generation (5G) wireless networks can be built upon the pillar of artificial intelligence concepts. After a general description of a cellular network and its management functionalities, highlighting the trends towards automatization, the paper focuses on the particular case of extracting knowledge about the time domain traffic pattern of the cells deployed by an operator. A general methodology for supervised classification of this traffic pattern is presented and it is particularized in two applicability use cases. The first use case addresses the reduction of energy consumption in the cellular network by automatically identifying cells that are candidates to be switched-off when they serve low traffic. The second use case focuses on the spectrum planning and identifies the cells whose capacity can be boosted through additional unlicensed spectrum. In both cases the outcomes of different classification tools are assessed. This capability to automatically classify cells according to some expert guidance is fundamental in future networks, where an operator deploys tenths of thousands of cells, so manual intervention of the expert is unfeasible.

Keywords: Classification. Cellular Networks. 5G. Radio Access Network Management.

1 Introduction

Our interconnected world is increasingly marked by fluid boundaries, tighter interlinkages and globally coordinated actions. Among these complexities, one of the most influential factors shaping our global society are networks. Networks serve as a central metaphor for describing the complexities of modern life. But they are also an undeniable technological foundation for unlocking tremendous social and economic benefits. Understanding the dynamics of networks – and their potential for positive change - can help us collectively meet our greatest social, economic and environmental challenges [1]. In this context, cellular networks have become pivotal: currently, there are as many mobile subscriptions as people in the world, and every second, 20 new mobile broadband subscriptions are activated. In addition to the increase in subscribers, data consumption also continues to rise.

Then, as a next step in the evolution of cellular communication systems, research carried out by industry and academia is nowadays focused on the development of the new generation of mobile and wireless systems, known as 5th Generation (5G) that targets a time horizon beyond 2020. 5G intends to provide solutions to the continuously increasing demand for mobile broadband services associated with the massive penetration of wireless equipment such as smartphones, tablets, the tremendous expected increase in the demand for wireless Machine To Machine communications [2], and the proliferation of bandwidth-intensive applications including high definition video, 3D, virtual reality, etc.

To cope with the abovementioned demands, requirements of future 5G system have been already identified and discussed at different fora [3][4]. Examples of these requirements include1000 times higher mobile data volume per area, 10 times to 100 times higher number of connected devices, 10 times to 100 times higher typical user data rate, 10 times longer battery life for low power devices and 5 times reduced End-to-End latency.

Furthermore, 5G networks will be fueled by the advent of big data and big data analytics [5]. The volume, variety and velocity of big data are simply overwhelming. Nowadays, there are already tools and platforms readily available to efficiently handle this big amount of data and turn it into value by gaining insight and understanding data structures and relationships, extracting exploitable knowledge and deriving successful decision-making. Applications of big data and big data analytics are already present in different sectors (e.g. entertainment, financial services industry, automotive industry, logistics, etc.). Therefore, with the huge amount of data generated by mobile networks, it can be envisaged that big data technologies will play a key role in 5G to extract the most of possible value of the available data exploiting it for enhancing the efficiency in mobile service provisioning.

In this context, this paper supports the idea that Artificial Intelligence (AI) mechanisms, which intend to develop intelligent systems able to perceive and analyze the environment and take the appropriate actions, will fully fertilize in the 5G ecosystem. While many seeds can be found in the literature both from an academic/theoretical perspective (e.g., connected to the so-called Cognitive Networks [6]) and from a practical perspective in current Third Generation (3G) and Fourth Generation (4G) networks (e.g., connected to the so-called Self-Organizing Networks [7]), more ambitious objectives can be targeted and the 5G era is the proper time for AI-based networks to happen.

Based on the above considerations, this paper intends to present a vision of how the different Radio Access Network (RAN) management procedures of future 5G networks can be built upon the pillar of AI concepts. For that purpose, the paper provides in Section 2 a general description of a cellular network and corresponding RAN management functionalities and highlights the trends towards automatization. This is mainly addressed to the non-specialized reader. In turn, Section 3 deepens on the framework to support RAN management in the context of 5G networks. Section 4 focuses on the particular case of extracting knowledge about the time domain traffic pattern of the cells and provides a number of potential applications. With this, two applicability use cases are presented in Section 5 and Section 6. Conclusions close the paper in Section 7.

2 Radio Access Network Management

A generic view of a cellular network is depicted in Fig. 1. Its main components are briefly described in the following: The User Equipment (UE) is the device that enables the Mobile Network Operator (MNO)'s customer to gain access to the network services (e.g., voice, data). The UE connects to the Radio Access Network (RAN) through the so-called radio interface (i.e., a wireless interface). Typically, the UE is multi-technology (e.g., 3G, 4G, WiFi) and can operate at different frequency bands.

The RAN is the network subsystem responsible to provide the connectivity between the UE and the Core Network (CN), which manages the provision of final services to the users. The RAN is composed of multiple base stations (BS), also known in general as "cells", and, for some technologies such as Second Generation (2G) and 3G, it also includes additional network controller nodes. The RAN includes a number of management functionalities in order to provide the wireless connectivity in an efficient way. In turn, the CN takes care of aspects such as the interconnection with other networks.

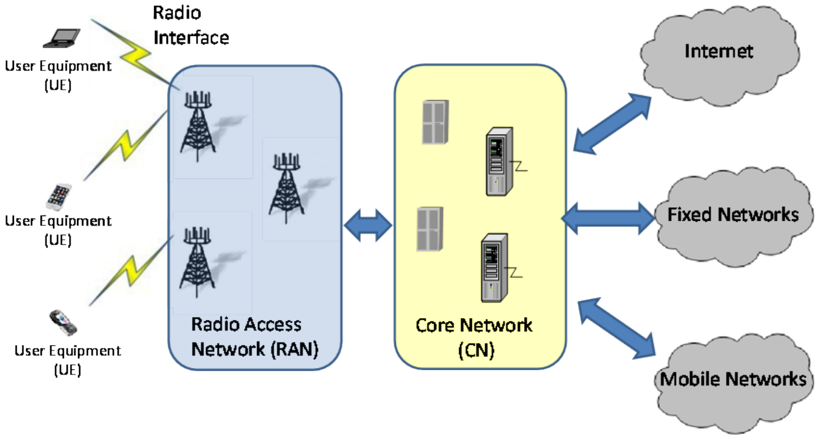

Fig. 1. Generic view of a cellular network architecture

Given that the RAN is composed of multiple cells and there will be multiple UEs moving around, as illustrated in Fig. 2, the set of RAN management functionalities takes care of deciding aspects such as to which cell a specific UE is attached to, the time at which a moving UE needs to switch the connectivity from one cell to another neighboring cell, how a given cell splits its capacity (i.e., data rate) among the different UEs that are attached to it, etc.

The vision of the future 5G RAN corresponds to a highly heterogeneous network at different levels, including multiple technologies, multiple cell layers, multiple spectrum bands, multiple types of devices and services, etc., with unprecedented requirements in terms of capacity, latency or data rates. The resulting network easily comprises 10.000-20.000 cells for a wide coverage service area (e.g., medium size European country). Consequently, the overall RAN management processes that constitute a key point for the success of the 5G concept will exhibit tremendous complexity. In this direction, legacy systems such as 2G/3G/4G already started the path towards a higher degree of automation in the planning and optimization processes through the introduction of Self-Organizing Network (SON) functionalities, in order to carry out these processes in a more efficient way.

SON refers to a set of features and capabilities designed to reduce or remove the need for manual activities in the lifecycle of the network, so that operating costs can be reduced as well as revenue can be protected by minimizing human errors. As such, with the introduction of SON features, classical manual planning, deployment, optimization and maintenance activities of the network can be replaced and/or supported by more autonomous and automated processes.

SON can greatly benefit from AI-based tools able to smartly process input data from the environment and come up with exploitable knowledge (i.e., knowledge that can be formalized in terms of models and/or structured metrics that represent the network behavior in a way that can be directly used to make smart network planning and optimization decisions). The obtained knowledge will drive the appropriate actions associated to the different SON functionalities. The target is to efficiently handle this big amount of data and turn it into value by gaining insight and understanding data structures and relationships, extracting exploitable knowledge and deriving successful decision-making.

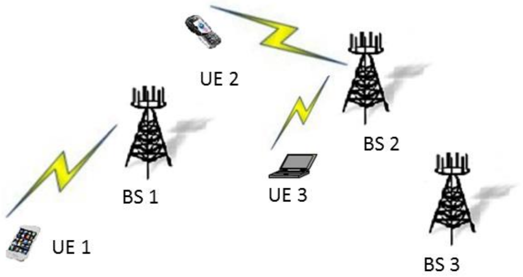

Fig. 2. Closer view of the RAN subsystem

3 Knowledge Discovery to support 5G RAN Management

3.1 Data Acquisition and Pre-processing

MNOs have traditionally operated with complex, disparate sets of data, with useful information residing in multiple systems such as customer relationship management systems, network management systems, billing, inventories, network elements, service management systems, deep packet inspection devices, application-specific databases, etc., [8]. In addition, MNOs have to deal with the concurrent operation of network elements belonging to multiple network generations (2G/3G/4G, etc.) and/or to multiple vendors, each one holding different types of data, in various formats. This huge heterogeneity of data and the associated difficulties in carrying out an efficient processing has led to perform the network management processes relying on a limited amount of data both in terms of variety of data considered (i.e., many counters and measurements that can be captured are not exploited at their possible extent) and time spam that are stored in the management systems (i.e., many counters and measurements are just retained for a short period of time in support of certain functionalities and then are either deleted or forgotten in back-up systems, while their applicability to keep the memory of the system is disregarded).

While this limited approach has been the rule in legacy 3G/4G systems, with SON deployment still at its infancy, a substantial evolution is necessary to deal with the increased complexity and stringent efficiency constraints of 5G. Therefore, the challenge for an efficient 5G RAN Management is to build this complete network vision by smartly analyzing and correlating all the different data sources in order to extract the most relevant information contained in them.

In general, gathered input data can belong to different categories (e.g., network data, user data, content data, external data). Network data characterizes the behavior of the network in terms of different measurements collected and recorded by the network nodes. Measurements include network traffic levels (e.g. traffic load at the radio interface, signaling traffic, active UEs per cell, etc.), resource access measurements, Quality of Service (QoS) measurements (e.g. throughput, latency, etc.) or cell availability measurements. Measurements can be performed by the network nodes (e.g. the cells) and also by the UEs that report them to the network.

The time span of the data collection will tightly depend on the targeted applicability of each type of data (e.g., planning actions will consider input data recorded over longer periods of time that can spam over several days or weeks while optimization actions will usually consider input data collected over much shorter time frames).

The proposed framework relies on the application of data mining techniques over the collected input data in order to distil all the available information and identify meaningful models and patterns that will drive the subsequent decisions. In this respect, the collected data coming from multiple heterogeneous sources needs to be preprocessed in order to prepare it for mining. This includes different tasks such as [9]: data cleaning to remove noise and inconsistent data (e.g. discard network counters that exhibit errors); data integration to combine network data collected at different nodes and exhibiting different time stamps or different periodicities; data selection to choose the relevant data for each specific analysis; and data transformation where data are consolidated through e.g. summary or aggregation operations (e.g. aggregating measurements collected with a periodicity of 15min to derive the equivalent measurement with 1h periodicity).

3.2 Knowledge Discovery

The Knowledge Discovery stage performs inference on the pre-processed data in order to build models that reflect the relevant knowledge that will drive the optimization and planning decisions. The core functionality of the knowledge discovery consists in learning from the users and the network in order to extract models that reveal their behavior. It is worth emphasizing here that, given the ultra-high level of efficiency that is associated to the design of future 5G systems, the target is to gain in-depth and detailed knowledge about the whole ecosystem, which in turn will enable ultra-efficient management and optimization. In this respect, the higher level of knowledge about the network and its users constitutes a key differential factor between 5G and legacy systems.

This stage will be based on machine learning tools used to carry out the mining of the pre-processed data to extract relevant knowledge at different levels: cell level (contains the characterization of the conditions on a per cell basis), cell cluster level (characterization of groups of cells built according to their similarities) and user level (contains the characterization of the existing conditions at the user equipment level).

The general goal of machine learning is to build computer systems that can adapt and learn from their experience [10]. Machine learning techniques are usually subdivided into three big categories, namely supervised learning, unsupervised learning and reinforcement learning. From the perspective of the knowledge discovery stage considered here, both supervised and unsupervised learning techniques are the ones that exhibit more applicability, while reinforcement learning tools will normally be more relevant for the decision making processes associated to the management functionalities.

Specific machine learning functions that are relevant in the framework considered here for RAN management are classification, prediction and clustering [9]. Among them, the focus of this paper is on the classification applied for knowledge discovery related to the time domain traffic variations of the cells deployed in a network. Classification is the process of finding a model or function that describes and distinguishes data classes or concepts. The obtained model (i.e. the classifier) is then used to determine the class to which an object belongs. The object to be classified is represented by a tuple that includes a set of attribute values. Classification process assumes that the possible classes are predefined in advance. Then, the classification model is usually obtained from a supervised learning algorithm that analyses a set of training tuples associated with known classes.

4 Classification of the cell-level time domain traffic

The cell-level time domain traffic defines how the traffic of a cell varies as a function of time. Traffic can be measured in different ways, such as the load factor, the total number of users connected to the cell, the total data rate, etc., and it can be aggregated or split among QoS classes. The traffic in a cell will be tightly related with the environment where the cell is deployed and with the characteristics and profiles of the users served by the cell. This will lead to time correlations in the traffic evolution of a given cell at different levels (e.g. intra-day variations in which the traffic can substantially differ between mornings or nights, variations during the week between working days and weekend, etc.). The detailed analysis of these correlations will allow extracting valuable knowledge that can be used for making management decisions regarding the configuration of a cell. In this respect, this paper focuses on the application of classification techniques to extract this knowledge. In particular, the cells will be classified based on their historical traffic samples. The possible classes will indicate certain behaviors of the cells that are relevant for different RAN management processes. In the following we start by providing the general classification methodology and then we particularize it according to its applicability in some selected use cases, providing some results obtained using data extracted from a real mobile network.

4.1 General classification methodology

The input data for each cell i is a time series $\mathbf{X_i}$=(x_i(t)$_,$, x_i(t-1), ..., x_i(t-(N-1))) composed of N samples of the measured traffic in the cell i at different times t. The objective of the classifier is to make an association between the input time series $\mathbf{X_i}$ and a class C($\mathbf{X_i}$) that characterizes the behavior of the cell's traffic in the time domain. The number and the type of classes will depend on the specific applicability of the classification outcomes, as it will be detailed in the use cases that will be presented later on.

Since the number of time samples N will typically be a very large value (e.g. reflecting the traffic measured in a cell in periods of some minutes and collected during several weeks, months, etc.), it will not be feasible to use the time series $\mathbf{X_i}$ directly as input of a classifier tool. Therefore, an initial processing is carried out to come up with a vector $\mathbf{F(X_i)}$ of shorter dimension M that preserves the relevant characteristics of the traffic pattern. This vector will be the input of the classifier. Following the usual terminology in classification [9], vector $\mathbf{F(X_i)}$ represents the tuple to be classified and each of its components represents a feature or attribute. Again, the definition of the mapping between $\mathbf{X_i}$ and $\mathbf{F(X_i)}$ will be dependent on the specific applicability of the classification, so it will be detailed later on when analyzing the different use cases.

The classifier will perform the association between the input $\mathbf{F(X_i)}$ and the class C($\mathbf{X_i}$), as illustrated in Fig. 3. The internal structure of the classifier will be given by the specific classification tool being used and its settings will be automatically configured through a supervised learning process executed during an initial training stage. This training will use as input S time series $\mathbf{X_j}$ j=1,...,S of some cells whose associated classes C($\mathbf{X}_j$) are pre-defined by an expert. In this way, the training set will be composed by the S tuples $\mathbf{F(X_j)}$, j=1,...,S and their associated classes C($\mathbf{X}_j$). The supervised learning process will analyze this training set to determine the appropriate configuration of the classification tool. The overall process is illustrated in Fig. 3.

4.2 Classification tools

Regarding the classification tool, the following alternatives are considered [9]:

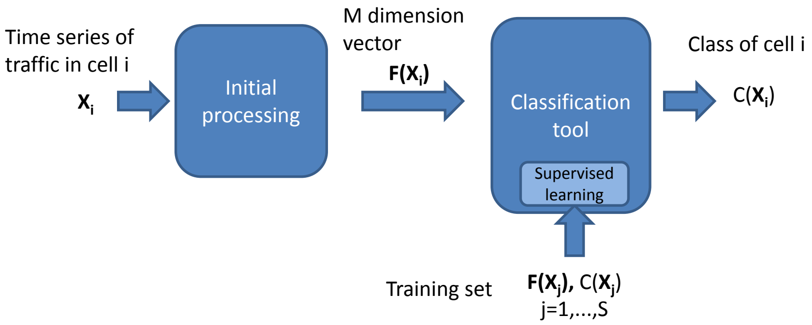

Fig. 3. General classification methodology

- Decision tree induction: The classification is done by means of a decision tree, which is a flow-chart structure where each node denotes a test on a feature value, i.e. a component of vector $\mathbf{F(X_i)}$, each branch represents an outcome of the test, and tree leaves represent the classes. The tree structure is built during the supervised learning stage through a top-down recursive divide-and-conquer manner, starting from the training set which is recursively partitioned into smaller subsets.
- Naive Bayes classifier: In this case the classifier evaluates the probability Prob(C($\mathbf{X_i}$)|$\mathbf{F(X_i)}$) that a given cell $\mathbf{X_i}$ belongs to a class C($\mathbf{X_i}$) based on the values of the features $\mathbf{F(X_i)}$. The resulting class is the one with the highest probability. The computation of this probability is done using Bayes' theorem under the "naive" assumption of class conditional independence, which presumes that the effect of a feature value on a given class is independent of the values of the other features. In turn, the different terms in the computation of the Bayes' theorem are obtained from the analysis of the training set.
- Support Vector Machine (SVM): A SVM is a classification algorithm based on obtaining, during the training stage, the optimal boundary that separates the vectors $\mathbf{F(X_j)}$ of the training set in their corresponding classes C($\mathbf{X_j}$). The obtained boundary is then used to perform the classification of any other input vector $\mathbf{F(X_i)}$. To find this optimal boundary, it uses a nonlinear mapping to transform the original training data into a higher dimension so that the optimal boundary becomes an hyperplane. Although SVM classifier is originally intended to do a binary classification, a multi-class SVM classifier can easily built by hierarchically combining multiple binary SVM classifiers. Each of these binary classifiers specifies whether the cell belongs or not to a given class.
- Neural Network: The classification is done by means of a feed-forward neural network that consists of an input layer, one or more hidden layers and an output layer. Each layer is made up of processing units called neurons. The inputs to the classifier, i.e. each of the components of vector $\mathbf{F(X_i)}$, are fed simultaneously into the neurons making up the input layer. These inputs pass through the input layer and are then weighted and fed simultaneously to a second layer. The process is repeated until reaching the output layer, whose neurons provide the selected class

C($\mathbf{X_i}$). The weights of the connections between neurons are learnt during the training phase using a back propagation algorithm.

The abovementioned general classification methodology presents applicability in different management process, such as network planning, optimization of radio resource management algorithms, energy saving, spectrum planning or load balancing. In the following sections, the methodology is particularized for two of these use cases.

5 Use case 1: energy saving

This use case aims at reducing the energy consumption in the deployed cellular network. According to the Mobile's Green Manifesto report [11], approximately 80% of the energy consumption and Green House Gas emissions of mobile operators is caused from their networks. From an economical perspective, if all networks with above-average energy consumption were improved to the industry average, there is a potential energy cost saving for mobile operators of $1 billion annually at 2010 prices. In case of improving to levels of the top quartile the cost saving could be more than $2 billion a year [11]. Therefore, techniques intended to reduce the energy consumption are relevant for operators of current and future networks.

In this use case, the energy reduction is done by switching off the cells that carry very little traffic at certain periods of the day (e.g. at night) and making the necessary adjustments in the neighbor cells so that the existing traffic can be served through some other cell. In this context, the classification methodology of section 4 can be used to identify candidate cells to be switched-off based on their traffic patterns. The automation of this procedure based on expert criteria captured in the training set becomes particularly useful considering that networks in the envisaged ultra-dense scenarios for future 5G systems can comprise several tens of thousands of cells. Therefore, it is not practical that a human expert can make this classification manually. It is worth mentioning, however, that the final decision on whether or not to switch off a cell would make use of this classification as well as other possible inputs which are out of the scope of this paper (e.g. the neighbor cell lists to ensure that a call that is generated in a cell that has been switched-off can be served through another cell).

In this use case a cell can be classified in two different classes:

- Class A: Candidate cell to be switched off
- Class B: Cell that cannot be switched off.

5.1 Data Acquisition and Pre-processing

In this case, the components of vector $\mathbf{F(X_i)}$ correspond to the average normalized traffic of the cell during the nights (i.e. from 0h to 8h), the mornings (i.e. from 8h to 16h) and the afternoons (i.e. from 16h to 24h) for each day of the week (Monday to Sunday). This leads to a total of M=21 components that can be easily obtained by normalizing the time series $\mathbf{X_i}$ so that the traffic ranges from 0 to 1 and by averaging the time series in each of the abovementioned periods.

To assess the behavior of the classification methodology in this use case, a set of real traffic measurements for a total of 419 cells deployed by an operator in a certain geographical region has been used. For each cell i, the time series $\mathbf{X_i}$ is composed by the data traffic measurements done every 15 min, and collected during a whole week. Therefore, each time series is composed by N=672 traffic samples. The traffic in a period of 15 min is given by the average number of users in the cell with an active data session.

5.2 Knowledge Discovery

The different classification tools discussed in section 4.1 have been implemented by means of RapidMiner Studio Basic [12]. The different parameters have been manually adjusted to obtain good accuracy levels of the different classification tools. In particular, the SVM is configured with radial kernel type, complexity constant which sets the tolerance for misclassification C=30, kernel cache 200 MB, convergence precision 0.001, a maximum of 10^5 iterations and the loss function is defined with complexity constants equal to 1 for both positive and negative examples and insensitivity constant equal to 0. The neural network classifier is configured with one hidden layer, 500 training cycles, learning rate 0.3, momentum 0.6 and the optimization is stopped if the training error gets below 10^{-5}. The decision tree is configured with maximal depth 20, minimal leaf size 2, confidence level 0.25, minimal size for split 4, minimal gain 0.1 and applying pruning and prepruning with 3 alternatives. Finally the Naive Bayes classifier is configured with Laplace correction, greedy estimation mode and 10 kernels.

5.3 Results

To illustrate the expert criteria to be learnt by the classification tool, Fig. 4 plots the time series $\mathbf{X_i}$ of 4 example cells included in the training set. Two of them are classified by the expert as A and two of them are classified as B. Then, different training sets have been built including these cells together with other examples in order to train the classification tools.

First, several tests have been done to derive the accuracy of the considered classification tools as a function of the training set size S. For a given S, the accuracy is measured by executing the classification over the cells of the training set and calculating the percentage of cells that are classified in the same category that was declared by the expert in the training. The test has been applied for all 4 classification tools and training set sizes ranging from S=10 to S=200. The best accuracy is obtained by the SVM, which provides 100% accuracy in all the cases, followed by the Neural Network and Decision Tree, which exhibit accuracy above 98.5%. The worst behavior is obtained with the Naive Bayes classifier with a minimum accuracy of 96.4%.

After completing the training process, the classification of the 419 available cells is performed. Then, as a first result that illustrates the operation of the classification process, Fig. 5 depicts the time series $\mathbf{X_i}$ of two example cells that didn’t belong to the training set: Cell 260, which is classified as Class A by all 4 classification tools considered, and Cell 240, which all 4 classification tools categorize as Class B. From visual inspection, and by comparing these cells with the examples given by the expert in Fig. 4, it appears an adequate decision given that Cell 260 exhibits relatively long periods at night serving no traffic at all and Cell 240 has traffic during all the time periods in the week.

Fig. 6 presents the total number of cells that are classified as A by each classification tool as a function of the training set size S. It is observed that, for low values of S (e.g. S=10) roughly half of the cells are classified as A and half are classified as B by all the tools. This indicates that, due to the low number of examples in the training set, the classification tools are not able to clearly distinguish the traffic patterns and the classification exhibits high randomness. Instead, when increasing the training set size S, the number of cells belonging to class A is substantially reduced for all the classifiers (e.g. for the case of the largest training set size S=200 the number of cells classified as A ranges from 46 with SVM up to 90 for the Naive Bayes case). It is worth emphasizing that the SVM exhibits a more efficient operation compared to the rest of classification tools since it is less sensitive to the value of S: as soon as the training set is S≥20, the result of the classification is very similar (i.e., there are around 50 cells classified as A).

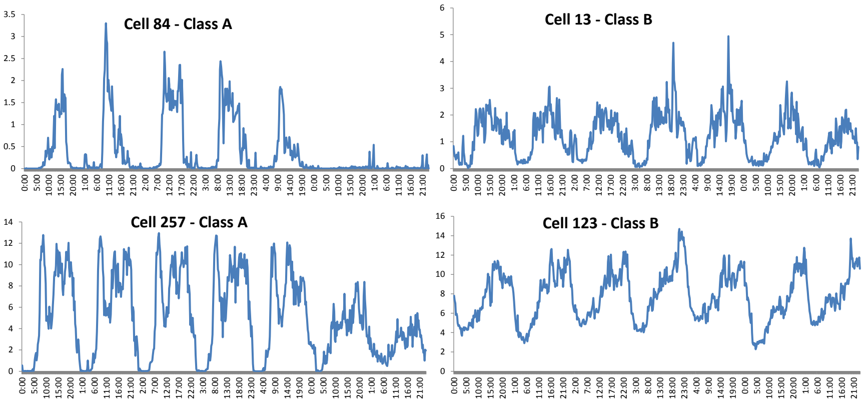

Fig. 4. Examples of cells of the training set belonging to classes A and B.

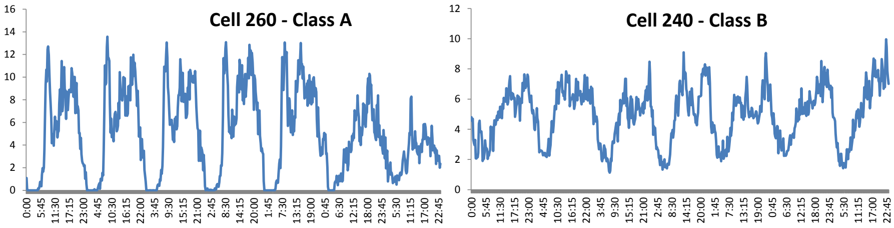

Fig. 5. Examples of two cells classified as A (Cell 260) and B (Cell 240).

Table 1 compares the outcomes of the different classification tools by presenting the percentage of coincidences between every pair of tools for the case S=200. For example, the table shows that 91% of the cells (i.e. 381 out of 419 cells) have been classified equally by the SVM and the Neural Network. The table also presents the "Expert validation", which measures the percentage of coincidences with respect to the classification made by the expert. It can be observed that the largest percentages of coincidences are obtained with SVM.

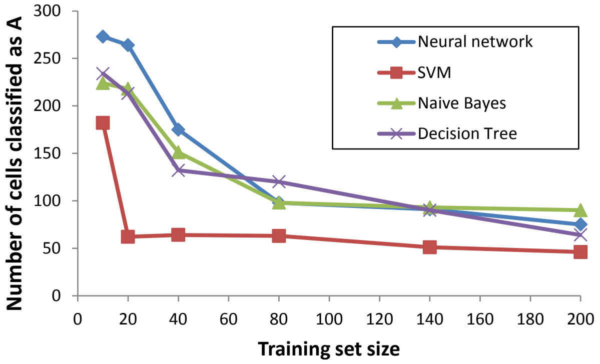

Fig. 6. Number of cells classified as A as a function of the training set size.

Table 1. Percentage of total coincidences by every pair of classification tools with S=200

	SVM	Neural Network	Naive Bayes	Decision Tree	Expert validation
SVM	--	91%	88%	93%	98%
Neural Network	91%	--	87%	91%	91%
Naive Bayes	88%	87%	--	88%	87%
Decision Tree	93%	91%	88%	--	94%

6 Use case 2: spectrum planning

In light of the more advanced spectrum management models envisioned for future 5G systems, the provisioning of the spectrum resources to be exploited at a given time and cell should be considered from a wider perspective. Specifically, although licensed spectrum remains operators' top priority to deliver advanced services and better user experience, other elements need to be explored as complements to meet the ultra-high capacity foreseen to be needed by future systems. These elements include the use of unlicensed spectrum considered in initiatives such as LTE-U (Unlicensed LTE) [13][14], as well as the use of shared spectrum on a primary/secondary basis in which the operator is allowed to access a certain spectrum band owned by a different primary user, as long as certain conditions are met in order not to interfere the primary users. With all these considerations, the use case considered here intends to decide whether it is possible or not to boost the capacity of a cell by exploiting unlicensed spectrum bands. This decision will exploit the knowledge about the time evolution of the cell's traffic, in the sense that typically unlicensed spectrum could be adequate to cope with sporadic traffic increases. Then, this use case intends to classify the cells according to the following classes:

- Class A: Candidate cell to boost capacity through additional unlicensed spectrum.
- Class B: Cell that does not need capacity boost through unlicensed spectrum.

6.1 Data Acquisition and Pre-processing

This use case has been assessed considering a total of 300 cells from a real cellular network deployed in an urban area, under the rationality that this type of scenario is where capacity boosting will be more likely needed. Besides, assuming that spectrum demands will be mainly associated to the periods of the day when there is more traffic, in this use case the components of vector $\mathbf{F(X_i)}$ correspond to the average traffic of a cell on a per hour basis, between 6h and 22h. This leads to a total of M=16 components. As a difference from the previous use case, here the traffic is not normalized, since the absolute value of the traffic is also relevant to decide whether additional unlicensed spectrum may be needed.

6.2 Knowledge Discovery

The same classification tools as in section 5.2 are considered here.

6.3 Results

Fig. 7 plots the components of vector of $\mathbf{F(X_i)}$ for 2 cells of the training set categorized as A and B by the expert. Class A cells use to exhibit peaks of high traffic levels while class B cells exhibit lower traffic values and more homogeneity. Like in the previous use case, different training set sizes have been used to train the considered classification tools. After the training process, the 300 cells have been classified. Fig. 8 depicts two example cells that were not included in the training set and that are classified as A and B by all the considered classifiers. It is observed that both cells present similar characteristics like the cells of the training set shown in Fig. 7, meaning that the classification tools have been able to identify also the relevant characteristics of the time evolution in this use case.

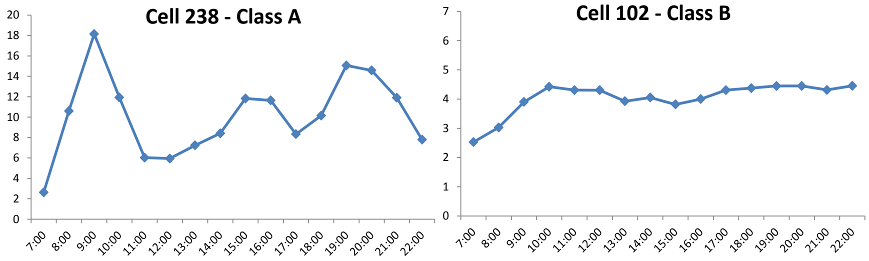

Fig. 7. Examples of cells of the training set belonging to classes A and B.

Fig. 9 presents the number of cells classified as A by each classifier as a function of the training set size with the different classifiers. Like in the previous use case it is observed that the SVM is able to converge more quickly than the other classifiers when the training set size is small. It is also noticed that for the case of S=140 very small differences are observed between the classifiers. This can also be corroborated in Table 2 that presents the percentage of coincidences between every pair of classifiers and with the expert validation. It can be observed that the percentages of coincidence with the expert in this use case are higher than in the previous one. This reflects that the characteristics that make a cell to be classified as A (e.g. sporadic traffic peaks) are more easily distinguishable than in the previous use case. Table 2 also shows that the best performance in terms of coincidences with the expert validation is achieved by both SVM and Neural Network classifiers.

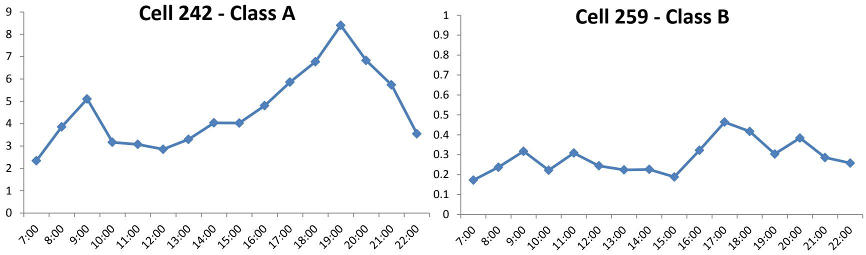

Fig. 8. Examples of two cells classified as A and B.

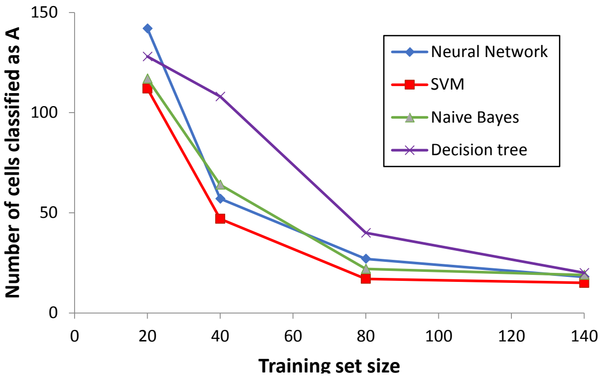

Fig. 9. Number of cells classified as A as a function of the training set size.

Table 2. Percentage of total coincidences by every pair of classification tools with S=140

	SVM	Neural Network	Naive Bayes	Decision Tree	Expert validation
SVM	--	97%	98.7%	97.7%	99.7%
Neural Network	97%	--	98.4%	99.4%	99.7%
Naive Bayes	98.7%	98.4%	--	99%	98.4%
Decision Tree	97.7%	99.4%	99%	--	97.4%

7 Conclusions

This paper has focused on the application of artificial intelligence and data mining concepts to support the radio access management in future cellular networks, where automatization is fundamental to cope with the huge number of cells that an operator can deploy, so manual intervention from a human expert becomes impractical. In particular, the paper has focused on extracting knowledge about the time domain traffic pattern of the cells. A general methodology for supervised classification of this traffic pattern has been presented and particularized in two applicability use cases, addressing energy saving and spectrum planning processes. In both cases the outcomes of different classification tools are assessed, concluding that the SVM technique is in general the one that best captures in the classification process the expert knowledge provided in the examples of the training set.

Acknowledgements. This work has been supported by the EU funded H2020 5G-PPP project SESAME under the grant agreement no 671596 and by the Spanish Research Council and FEDER funds under RAMSES grant (ref. TEC2013-41698-R).

References

1. World Economic Forum, "Enabling Transformation: Information and Communications Technologies and the Networked Society", 2009.
2. Ericsson White Paper "More than 50 billion connected devices", February, 2011, http://www.ericsson.com/res/docs/whitepapers/wp-50-billions.pdf
3. M. Fallgren, B. Timus, (editors), "Scenarios, requirements and KPIs for 5G mobile and wireless system", Deliverable D1.1. of the METIS project, May, 2013.
4. R. El Hattachi, J. Erfanian (editors) "NGMN 5G White Paper", NGMN Alliance, February, 2015
5. Ericsson, "Big Data Analytics", White paper, August, 2013.
6. R.W. Thomas, L.A. DaSilva, A.B. MacKenzie, "Cognitive Networks", 1st IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, (DySPAN), 2005, pp.352-360.
7. J. Ramiro, K. Hamied, Self-Organizing Networks. Self-planning, self-optimization and self-healing for GSM, UMTS and LTE, John Wiley & Sons, 2012.
8. A. Banerjee, "Advanced Predictive Network Analytics: Optimize your Network Investments and Transform Customer Experience", White Paper, Heavy Reading, February, 2014.
9. J. Han, M. Kamber, "Data Mining Concepts and Techniques", 2nd edition, Elsevier, 2006.
10. R.A.Wilson, F.C.Keil, The MIT Encyclopedia of the Cognitive Sciences, MIT Press, 1999.
11. GSMA, Mobile's Green Manifesto, 2nd Edition, June 2012.
12. RapidMiner Studio, http://www.rapidminer.com
13. 3GPP workshop on LTE in unlicensed spectrum, Sophia Antipolis, France, June 13, 2014. http://www.3gpp.org/ftp/workshop/2014-06-13_LTE-U/
14. A. Al-Dulaimi, S. Al-Rubaye, N. Quiang, E. Sousa, "5G Communications Race: Pursuit of More Capacity Triggers LTE in Unlincensed Band", IEEE Vehicular Technology Magazine, Vol. 10, Issue 1, pp. 43-51, February 2015, DOI: 10.1109/MVT.2014.2380631