DECENTRALIZED LOAD BALANCING IN MOBILE COMMUNICATION NETWORKS

Florian Bahlke and Marius Pesavento

Communication Systems Group, Technische Universität Darmstadt, Germany

ABSTRACT

Future generations of mobile communication networks are envisioned to utilize dense deployments of heterogeneous cell types to fulfill increasing performance requirements. An efficient and optimized utilization of the network resources, including user allocation management and load balancing between cells, is crucial for such networks to maintain high throughput and to handle increasing interferences. Due to the combinatorial nature of user allocation, load balancing is a mixed-integer linear problem that does not scale well will the number of users and cells. Heuristic methods to solve similar problems in this context are available, but they typically require extensive coordination between network entities while achieving highly suboptimal performance. We propose a machine learning based approach that achieves close to optimal performance while requiring very limited local interaction and computational effort during operation.

Index Terms— mobile communications, heterogeneous networks, load balancing, network optimization, support vector machines

1. INTRODUCTION

For the upcoming fifth generation of mobile communication networks (5G), multiple technologies are subject of current research. With the use of state-of-the-art modulation and coding schemes, the throughput of a conventional networks is being pushed to its theoretical limit. To satisfy increasing performance requirements, the network needs to utilize additional resources, which can be of various types [1, 2, 3, 4]. Research in millimeter-wave communications focuses on utilizing additional frequency bands in the wireless spectrum, while Massive-MIMO employs large antenna arrays to achieve a spatially orthogonalized transmission and reception of signals. Supplementing these developments in the wireless network architecture is the dense deployment of heterogeneous networks, or in short "HetNets" [5, 6].

The concept of HetNets has already matured with the current fourth generation of mobile communication networks, but will only reach truly widespread and massive deployment in upcoming generations [7]. Macro cells (MC) with large coverage areas are assisted in serving user demands by multiple small cells (SC), which have lower transmit power and therefore lower coverage areas. These small cells form an additional tier in the network, which has shown to be very suitable for assisting the service of multi-user hotspots. A key challenge however is to analyze and balance the load experienced by both tiers, the macro- and the small cells, such that the overloading of single cells and underutilization of others is prevented [8, 9, 10]. Established approaches to balance the cell load by optimizing user allocation typically require solving optimization problems that rely on global knowledge of the statuses of multiple entities in the network [11, 12]. Decentralized schemes have been proposed that implicitly achieve load balancing by artificially expanding the coverage areas of the (typically underutilized) small cells in a process referred to as "cell range expansion". Decentralized methods for cell range expansion have been developed [13, 14], and proven to show good performance. Close to optimal performance however can only be achieved by maintaining and updating tables of system parameters according to changes in the network. This process requires continuous adaptation and information exchange between cells, which has been identified as potential bottlenecks [15, 16].

We introduce an approach to solve the load balancing problem optimally based on Mixed-Integer Linear Programming (MILP). Because this problem is combinatorial in nature it does not scale well for high numbers of users and cells. Also, the problem needs to be solved jointly for, and using information from, all entities in the network. Because of these impracticalities we will use the MILP-based problem formulation only as a performance benchmark and for providing training data to our machine learning based method. This decentralized method provides resource allocation by utilizing multi-class support vector machines that traditionally are used for classification problems. Each user is allocated based on information available directly from only those neighboring cells that are potential allocation candidates. We define, without loss of generality, the "primary", "secondary" and "tertiary" allocation candidates under consideration as those cells that provide, in decreasing order of magnitude, the strongest signal to the user. The information provided by the cells to support the balancing procedure includes their type (macro- or small cell), an estimation of the current data demand from users in their coverage area, and information about the channel conditions of the users. These attributes are then used in a scheme that relies on support vector machines (SVM) to

"classify" each user as belonging to any of the three candidate cells. As training data for the SVM, we use the results from the optimization problem of minimizing the maximum load over the entire network. Once the SVM is trained however, the obtained learning system can be used in a decentralized way. Subsequent optimization is only necessary when the cell topology significantly changes, for example if cells are being shut off entirely. Preventing extensive optimization and continuous information exchange is the key benefit of the proposed method compared to established approaches, i.e. [13, 8, 12]. As an additional requirement compared to said methods, our approach requires data for training the support vector machines, but historic data is generally available and the costly MILP-based network optimization to obtain training labels can be carried out during the design phase. The application of statistical learning methods in optimizing wireless communications networks is only being considered recently [17], and to the best of our knowledge, methods similar to the proposed one have not been previously introduced. The remainder of the paper is organized as follows: In Sec. 2, we provide a system model for the wireless communication network, followed by a discussion of the proposed scheme in Sec. 3. Simulation results and a comparison of the proposed method with the conventional approach, as well as the optimal solution, are given in Sec. 4, and a summary of the work is provided in Sec. 5.

Notation: Throughout the paper, we will use normal letters for scalars, bold lowercase letters for column vectors and bold uppercase letters for matrices. We further indicate with $||\cdot||$ the euclidean norm of a vector, and with $\cdot^T$ the vector transpose.

2. SYSTEM MODEL

Let us consider a wireless communication network with k cells for $k = 1, \ldots, K$. We define a cell as the coverage area of one base station antenna, which has the transmit power p_k. It is further assumed that the network contains M demand points (DP), indicated by $m = 1, \ldots, M$. The parameter D_m of each DP may represent the demand of single users, or aggregated data demand of hotspots where all user nodes are in close proximity, and that therefore can be assumed to have similar channel conditions. The attenuation factor between the antenna of cell k and DP m is the product of path loss and antenna gain attenuation factors, and will in the following be denoted as g_{km}. The Signal-to-Interference-plus-Noise ratio (SINR) of cell k serving user m can be computed as

$$\gamma_{km} = \frac{p_k g_{km}}{\sum_{j=1, j\neq k}^{K} p_j g_{jm} + \sigma^2} \tag{1}$$

where σ^2 represents power of additive white Gaussian noise. We denote W as the total available system bandwidth and as η_{km}^{BW} the bandwidth efficiency of the communication link between cell k and DP m. To satisfy the demand

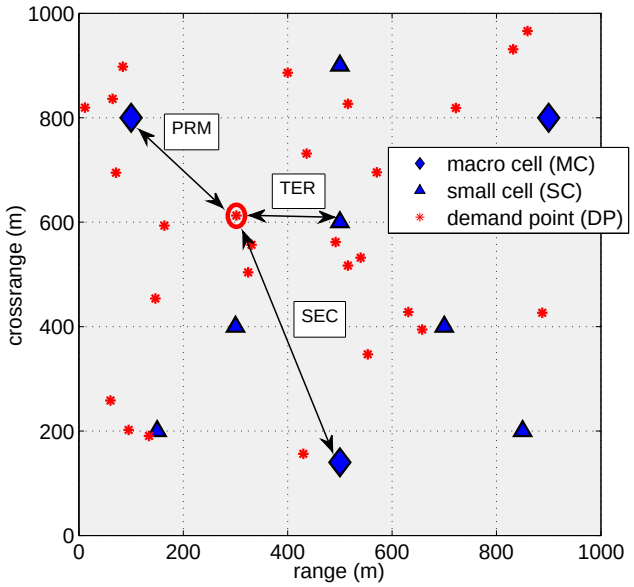

Fig. 1. network scenario and allocation example

of DP m, cell k needs to use the fraction $\Phi(k,m) = D_m / \left(\eta_{km}^{\mathrm{BW}} W \log_2(1+\gamma_{km})\right)$ of its resources [18, 19]. We define the matrix $\boldsymbol{A} \in \{0,1\}^{K\times M}$ with its elements A_{km} set as $A_{km} = 1$ if DP m is allocated to cell k, and $A_{km} = 0$ otherwise. Therefore the total ratio of its resources cell k is utilizing to satisfy the data demand of allocated users can be computed as

$$\rho_k(\boldsymbol{A}) = \sum_{m=1}^{M} A_{km} \frac{D_m}{\eta_{km}^{\mathrm{BW}} W \log_2(1+\gamma_{km})}. \tag{2}$$

For feasible network scenarios, where no cell is overloaded, it holds that $0 \leq \rho_k \leq 1 \; \forall k$. The maximum load level of any cell in the network is $\max_k \rho_k$.

For each DP m, we consider three neighboring cells as candidates for the allocation, which is illustrated in Fig. 1. The primary, secondary and tertiary allocation candidates of DP m are, in descending order of magnitude, those cells which can provide the first-, second-, and third-highest signal power $p_k g_{km}$ at the DP's location. Their indices are listed in the vectors $\boldsymbol{\kappa}^{\mathrm{P}}, \boldsymbol{\kappa}^{\mathrm{S}}, \boldsymbol{\kappa}^{\mathrm{T}} \in \{0,1\}^{M\times 1}$, respectively, with their respective elements determined as

$$\kappa_m^{\mathrm{P}} = \arg\min_k p_k g_{km}, \tag{3}$$

$$\kappa_m^{\mathrm{S}} = \arg\min_{k\setminus\{\kappa_m^{\mathrm{P}}\}} p_k g_{km}, \tag{4}$$

$$\kappa_m^{\mathrm{T}} = \arg\min_{k\setminus\{\kappa_m^{\mathrm{P}}, \kappa_m^{\mathrm{S}}\}} p_k g_{km}. \tag{5}$$

In the case that we set $A_{km} = 1$ for $k = \kappa_m^{\mathrm{P}}$ for all DPs m, each DP is allocated according to the cell providing the connection with the highest SINR. In this way the additional load

caused by each individual connection is minimized. We will employ this common allocation scheme as a heuristic baseline approach, in the following denoted as "max. SINR".

3. LOAD BALANCING

The load balancing problem in a wireless communication network can be solved optimally for the full network, by choosing the allocation of DPs to cells such that the maximum load level α among all cells in the network is minimized. The corresponding mixed-integer linear optimization problem is the following:

$$\underset{\alpha, \boldsymbol{A}^*}{\text{minimize}} \quad \alpha \tag{6a}$$

$$\text{subject to} \quad \alpha \geq \sum_{m=1}^{M} A^*_{km} \Phi(k,m) \; \forall k \tag{6b}$$

$$\sum_{k=1}^{K} A^*_{km} = 1 \; \forall m \tag{6c}$$

$$\alpha \in \mathbb{R}_{0+} \tag{6d}$$

$$A^*_{km} \in \{0,1\} \; \forall k,m \tag{6e}$$

In problem (6), Eq. (6c) constrains each DP to be allocated to exactly one cell. As discussed in Sec.1, we intend to use this approach as a benchmark and to provide training data to the machine learning based scheme as introduced in the following.

Let us define the label vector $\boldsymbol{y} \in \mathbb{N}^{M\times 1}$ which has elements y_m determined as follows:

$$y_m = \begin{cases} 2 \text{ if } A^*_{\kappa^{\mathrm{S}}_m m} = 1 \\ 3 \text{ if } A^*_{\kappa^{\mathrm{T}}_m m} = 1 \\ 1 \text{ otherwise} \end{cases} \tag{7}$$

To obtain training datasets of sufficient size, we use the labels of all DPs from N tests for a total of $T = MN$ labels, indicated by $t = 1, \ldots, T$. We will therefore refer to the t-th label as y_t.

For the training of the proposed statistical learning approach for user allocation, attributes need to be extracted for the three candidate allocation cells of each user m. These attributes are designed to reflect our specific knowledge of the network and the parameters we find significant for the allocation problem. We extract three attributes for each cell. The first attribute is an indicator of cell type defined as

$$\Lambda(k) = \begin{cases} 1 & \text{if cell } k \text{ is a small cell,} \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

The second attribute describes the additional load that user m would cause to cell k if it was allocated to it. This corresponds exactly to the parameter $\Phi(k,m)$ introduced in Sec.2.

The third attribute is the "would-be" load of cell k, if the max. SINR scheme was being used. This attribute serves as a measure of load caused by DPs in each cell's coverage area. The third attribute is determined as follows:

$$\Psi(k) = \sum_{m|\kappa^{\mathrm{P}}_m = k} \Phi(k,m). \tag{9}$$

Using all three of the aforementioned attributes for each of the three candidate cells for allocation, we can determine the attribute vector of DP m as

$$\boldsymbol{x}_m = \big[\Lambda(\kappa^{\mathrm{P}}_m), \Lambda(\kappa^{\mathrm{S}}_m), \Lambda(\kappa^{\mathrm{T}}_m), \Phi(\kappa^{\mathrm{P}}_m, m), \Phi(\kappa^{\mathrm{S}}_m, m), \ldots \\ \Phi(\kappa^{\mathrm{T}}_m, m), \Psi(\kappa^{\mathrm{P}}_m), \Psi(\kappa^{\mathrm{S}}_m), \Psi(\kappa^{\mathrm{T}}_m)\big]^T. \tag{10}$$

Similar to the label vector, we aggregate attribute vectors from N tests for a total of $T = NM$ attribute vectors, and we will therefore refer to the t-th attribute vector sample as $\boldsymbol{x}_t$.

This results in a training problem of a multi-class classifier which we solve by training two SVM, with the first being used to identify DPs that are allocated to their secondary cell, and the second SVM identifying those that are allocated to their tertiary cell, characterized by the normal vectors to their separating hyperplanes $\boldsymbol{w}^{21}$ and $\boldsymbol{w}^{31}$ respectively. The optimization problem to be solved for training an SVM that classifies between classes i and j is the following [20]:

$$\underset{\boldsymbol{w}^{ij}, b^{ij}, \boldsymbol{\xi}^{ij}}{\text{minimize}} \quad \frac{1}{2}(\boldsymbol{w}^{ij})^T \boldsymbol{w}^{ij} + C \sum_t \xi^{ij}_t \tag{11a}$$

$$\text{subject to} \quad (\boldsymbol{w}^{ij})^T \phi(\boldsymbol{x}_t) + b^{ij} \geq 1 - \xi^{ij}_t \text{ if } y_t = i \tag{11b}$$

$$(\boldsymbol{w}^{ij})^T \phi(\boldsymbol{x}_t) + b^{ij} \leq \xi^{ij}_t - 1 \text{ if } y_t = j \tag{11c}$$

$$\xi^{ij}_t \geq 0 \tag{11d}$$

$$\boldsymbol{w}^{ij} \in \mathbb{R}^{L\times 1}, b^{ij} \in \mathbb{R}, \boldsymbol{\xi}^{ij} \in \mathbb{R}^{L\times 1} \tag{11e}$$

In problem (11), the function $\phi(\boldsymbol{x}_t)$ maps the 9-dimensional attribute vector $\boldsymbol{x}_t$ onto the L-dimensional feature space, where for example polynomial combinations of the attributes are also being considered. SVM training problems such as (11) can be solved efficiently in their dual formulation using kernel functions [21], a functionality which is included in common machine learning software tools [22, 23].

We train two SVMs to obtain $\boldsymbol{w}^{21}$ and $\boldsymbol{w}^{31}$. We define $\hat{y}_m$ as the cell type that is classified by the SVMs according to the two decision functions, which is computed as:

$$\hat{y}_t = \begin{cases} 2 \text{ if } (\boldsymbol{w}^{21})^T \phi(\boldsymbol{x}_m) + b^{21} \geq 0 \text{ and} \\ \quad (\boldsymbol{w}^{21})^T \phi(\boldsymbol{x}_m) + b^{21} \geq (\boldsymbol{w}^{31})^T \phi(\boldsymbol{x}_m) + b^{31} \\ 3 \text{ if } (\boldsymbol{w}^{31})^T \phi(\boldsymbol{x}_m) + b^{31} \geq 0 \text{ and} \\ \quad (\boldsymbol{w}^{31})^T \phi(\boldsymbol{x}_m) + b^{31} \geq (\boldsymbol{w}^{21})^T \phi(\boldsymbol{x}_m) + b^{21} \\ 1 \text{ otherwise} \end{cases} \tag{12}$$

Using Eq. (12), we obtain the allocation decisions for all DPs m in a given network scenario, which lead to a load-balanced allocation solution for the full network.

4. SIMULATION RESULTS

Simulations of a wireless communication network with three macro- and six small cells in fixed positions on a 1000×1000m area as illustrated in Fig. 1 are carried out. The macro cells have a transmit power of 46 dBm and 10 dB antenna gain, totaling an equivalent effective isotropic radiated power (EIRP) of 56 dBm. All small cells are simulated with 36 dBm transmit power and 5 dB antenna gain, and therefore exhibit an EIRP of 41 dBm. Path loss is simulated according to the specifications in 3GPP TS 36.814 [24]. The bandwidth efficiency is set to $\eta_{km} = 0.8$ and the total system bandwidth is $W = 20$MHz.

Problem (6) is solved using the CVX toolbox for MATLAB [25] with the Gurobi solver [26], and the SVM training problem (11) is solved using the Machine Learning Toolbox for Matlab [23]. For the training of the SVMs, we use $T = 10000$ DP attribute vectors from $N = 100$ simulations of network scenarios with $M = 100$ DPs each. The soft threshold weighting parameter C in problem (11) is determined by searching on a grid the value that provides the highest classification accuracy on the training set. For the function $\phi(\cdot)$ in problem (11) consider both, the linear mapping of attributes to features, and the mapping to quadratic features. In the following we refer to these two methods "lin. SVM" and "quad. SVM", respectively. To evaluate the performance of the algorithms with a testing set, we generate $N = 100$ instances of network scenarios with $M = 100$ DPs each and average the resulting load levels of each cell over all scenarios. Accordingly, SVM classification on the testing set is performed using the coefficients obtained from the training set.

As observable in Fig. 2, the maximum cell load increases with the demand, here simulated in the range of 0-1 MBit/s per DP. For the given network configuration, there are differences in the sizes of coverage areas that result in unbalanced load levels across the cells. The max. SINR approach is not designed to mitigate this imbalance, and therefore exhibits the worst performance. Both SVM-based methods however perform better than the max. SINR approach, with the quadratic SVM being very close to the optimal solution. This demonstrates that using the learning-based approach discussed in the paper, we can obtain a decentralized load balancing scheme that is close to the globally optimal solution of a computationally extensive, joint optimization of all allocations in the network.

The average load level of individual cells for all methods and a fixed DP demand of 1 MBit/s is shown in Fig. 3. It shows that the cell "MC1", which corresponds to the macro cell in the lower center area of Fig. 1, is close to being overloaded. Both the SVM-based methods and the optimal solution achieve this through offloading to small cell. It is observable that the small gap to the optimal solution probably originated from small cell "SC2" being underutilized in the SVM-based methods compared to the optimal solution.

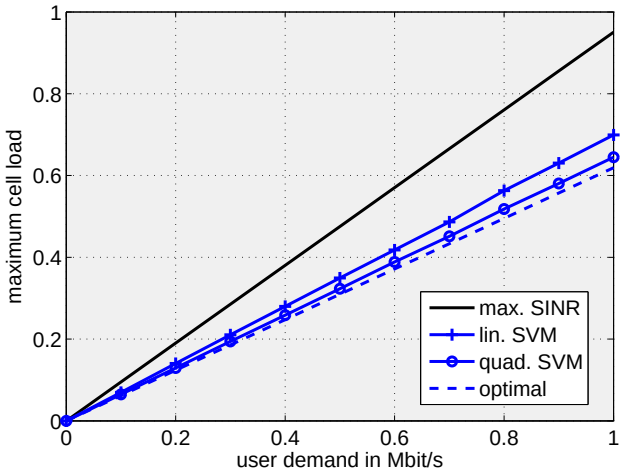

Fig. 2. maximum load comparison of allocation schemes over user demand

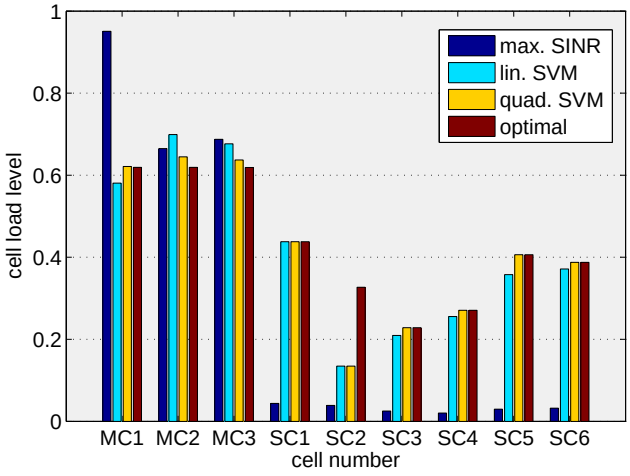

Fig. 3. cell load comparison of allocation schemes for fixed user demand

5. CONCLUSION

We introduced a method for decentralized load balancing in mobile communication networks that relies on using a support vector machines based classification procedure for resource allocation. Compared to many established approaches, this method does not require joint optimization of multiple network entities, nor does it create high communication overhead, and during operation it only requires limited local interaction. Simulation results reveal that the proposed method provides performance that is very close to the optimum. Further research can be dedicated towards a method that additionally adapts to significant changes in the network topology. The SVM coefficients could be automatically updated, for example if cells deactivate for power saving.

6. REFERENCES

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What Will 5G Be?," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[2] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five Disruptive Technology Directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, February 2014.

[3] M. Iwamura, "NGMN View on 5G Architecture," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015, pp. 1–5.

[4] Y. Cheng, M. Pesavento, and A. Philipp, "Joint Network Optimization and Downlink Beamforming for CoMP Transmissions Using Mixed Integer Conic Programming," *IEEE Transactions on Signal Processing*, vol. 61, no. 16, pp. 3972–3987, Aug 2013.

[5] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer, "Network Densification: The Dominant Theme for Wireless Evolution into 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, February 2014.

[6] X. Ge, S. Tu, G. Mao, C. X. Wang, and T. Han, "5G Ultra-Dense Cellular Networks," *IEEE Wireless Communications*, vol. 23, no. 1, pp. 72–79, February 2016.

[7] L. Wang, K. K. Wong, R. W. Heath, J. Yuan, and J. Yuan, "Wireless Powered Dense Cellular Networks: How Many Small Cells Do We Need?," *IEEE Journal on Selected Areas in Communications*, vol. PP, no. 99, pp. 1–1, 2017.

[8] I. Siomina and D. Yuan, "Load Balancing in Heterogeneous LTE: Range Optimization via Cell Offset and Load-Coupling Characterization," in *2012 IEEE International Conference on Communications (ICC)*, June 2012, pp. 1357–1361.

[9] F. Bahlke, O. D. Ramos-Cantor, and M. Pesavento, "Budget Constrained Small Cell Deployment Planning for Heterogeneous LTE Networks," in *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, June 2015, pp. 1–5.

[10] Z. H. Yang, Y. J. Pan, M. Chen, H. Xu, and J. F. Shi, "Cell Load Coupling with Power Control for LTE Network Planning," in *2015 International Conference on Wireless Communications Signal Processing (WCSP)*, Oct 2015, pp. 1–5.

[11] L. You and D. Yuan, "Load Optimization With User Association in Cooperative and Load-Coupled LTE Networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 3218–3231, May 2017.

[12] Q. Kuang and W. Utschick, "Energy Management in Heterogeneous Networks With Cell Activation, User Association, and Interference Coordination," *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 3868–3879, June 2016.

[13] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User Association for Load Balancing in Heterogeneous Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, June 2013.

[14] M. Shirakabe, A. Morimoto, and N. Miki, "Performance Evaluation of Inter-Cell Interference Coordination and Cell Range Expansion in Heterogeneous Networks for LTE-Advanced Downlink," in *2011 8th International Symposium on Wireless Communication Systems*, Nov 2011, pp. 844–848.

[15] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An Overview of Load Balancing in HetNets: Old Myths and Open Problems," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 18–25, April 2014.

[16] A. Damnjanovic, J. Montojo, Yongbin Wei, Tingfang Ji, Tao Luo, M. Vajapeyam, Taesang Yoo, Osok Song, and D. Malladi, "A Survey on 3GPP Heterogeneous Networks," *Wireless Communications, IEEE*, vol. 18, no. 3, pp. 10–21, June 2011.

[17] C. Jiang, H. Zhang, Y. Ren, Z. Han, K. C. Chen, and L. Hanzo, "Machine Learning Paradigms for Next-Generation Wireless Networks," *IEEE Wireless Communications*, vol. 24, no. 2, pp. 98–105, April 2017.

[18] K. Majewski and M. Koonert, "Conservative Cell Load Approximation for Radio Networks with Shannon Channels and its Application to LTE Network Planning," in *2010 Sixth Advanced International Conference on Telecommunications (AICT)*, May 2010, pp. 219–225.

[19] I. Siomina and Di Yuan, "Analysis of Cell Load Coupling for LTE Network Planning and Optimization," *IEEE Transactions on Wireless Communications*, vol. 11, no. 6, pp. 2287–2297, June 2012.

[20] Chih-Wei Hsu and Chih-Jen Lin, "A Comparison of Methods for Multiclass Support Vector Machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, Mar 2002.

[21] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An Introduction to Kernel-Based Learning Algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, Mar 2001.

[22] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[23] "MATLAB and Statistics Toolbox Release 2013a, The MathWorks, inc., Natick, Massachusetts, United States," .

[24] "3GPP TS 36.814 Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects v9.0.0," .

[25] Michael Grant and Stephen Boyd, "CVX: Matlab Software for Disciplined Convex Programming, version 2.1," `http://cvxr.com/cvx`, Mar. 2014.

[26] "Gurobi Optimizer 6.0," `www.gurobi.com`.