Azimuth Estimation for Sectorized Base Station With Improved Soft-Margin Classification

CHENAO WENG\textsuperscript{1}, HAI WANG\textsuperscript{2}, KE LI\textsuperscript{1}, AND M. N. S. SWAMY\textsuperscript{3}, (Life Fellow, IEEE)

\textsuperscript{1}College of Smart City, Beijing Union University, Beijing 100101, China
\textsuperscript{2}School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
\textsuperscript{3}Department of Electrical and Computer Engineering, Concordia University, Montreal, QC H3G 1M8, Canada

Corresponding author: Ke Li (like@bnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61972040, in part by the Collaborative Innovation Project of Chaoyang District of Beijing under Grant CYXC1805, and in part by the Premium Funding Project for the Academic Human Resources Development in Beijing Union University under Grant BPHR2018CZ05.

\textbf{ABSTRACT} Sector antenna azimuth plays an important role in the operation and maintenance of mobile networks. It is adjusted frequently so as to guarantee the high quality of coverage and low interference among neighboring cells. As one of the key elements of base station almanac (BSA), the azimuth of all the base stations need to be precisely managed by the operator. However, currently it is generally acquired by on-site measurement and updated manually, which is neither timely nor cost-effective. In addition, it is not open for third parties who need the information for network analysis. In view of this, by transforming the problem of azimuth estimation into one of searching for the optimal cell boundaries under the constraint of site location, this paper proposes an azimuth estimation method based on an improved multi-class soft-margin support vector machine (SVM). The explicit expression of the objective function of the problem is deduced through Lagrange duality. A circular one-versus-one strategy is utilized in dealing with the multi-class classification problem. In addition, a so-called average confusion index is designed to evaluate the degree of boundary separability of a base station so as to prejudge whether it is worthwhile to estimate its azimuths. Experiments are undertaken to validate the proposed algorithm by utilizing the spatial-temporal wireless signal dataset crowdsensed from massive user terminals in the 4G network. The experiments show that the proposed algorithm gives a significantly higher accuracy compared to that of two existing methods, namely the Gauss and radial rasterization methods. The impacts of data volume, penalty term and the boundary separability on the azimuth estimation is also discussed. The new algorithm is a promising alternative to the conventional manual solution with a much higher efficiency and lower cost, thus improving the intelligence of mobile network operation.

\textbf{INDEX TERMS} Mobile crowdsensing, base station almanac, soft-margin SVM, sector azimuth.

\section{I. INTRODUCTION}

For mobile networks, cells are the basic unit in network operation. As the core database for network operation and maintenance, base station almanac (BSA) describes the fundamental parameters of all the cells or sectors in a network. It is highly confidential and generally difficult to obtain. Meanwhile, with the expansion and optimization of the network, there are always new cells deployed, and existing cells removed or relocated. In routine network optimization and operations, the antenna azimuth and downtilt of a sectorized base station (BS) are adjusted frequently to improve its coverage. Therefore, BSA is dynamically changed, and it is difficult even for the carriers themselves to maintain the BSA information timely and accurately.

The optimal setting of antenna azimuth is very important for the quality of mobile network coverage. On the one hand, precise setting of the azimuth can ensure that the actual coverage of the sectorized base station is the same as planned, and ensure the quality of network coverage. On the other hand, adapting the azimuth timely according to the traffic distribution can better optimize the network coverage capability.

At present, the acquisition of antenna parameters depends mainly on the on-site measurement with compass,
inclinometer and other instruments. The accuracy, efficiency, and timeliness of the measurements are greatly affected by artificial factors. With the expansion of the network scale, the manual measurement method has greatly affected the efficiency and quality of network maintenance and optimization, and it is urgent to acquire and manage it in an automated manner, thereby improving the work efficiency and reducing the maintenance and optimization costs. In addition, it is not open for third-parties who need the information for network analysis and location-based services.

In view of the above problems, this paper proposes a machine learning based method for detecting the sector azimuth with higher efficiency and lower cost. By utilizing the coverage information collected on a large number of user terminals, an improved multi-class soft-margin SVM algorithm is employed to estimate the boundary hyperplane of neighboring cells of the base station. In addition to this, judgement of the direction of the boundary vector is also addressed and then the complete method of sector azimuth estimation is presented. To improve the practicability of the proposed method, a mechanism, namely average confusion index (ACI), is designed to preclude the base stations with severe spatial sample confusion.

The remainder of the paper is as follows: Chapter II introduces related work, including BSA and its construction, sector antenna azimuth detection, SVM and the dataset utilized in this article. Chapter III proposes a sector azimuth estimation method based on an improved soft-margin SVM. Chapter IV verifies the effectiveness of the proposed method based on the data collected in the real 4G network, and compares it with traditional methods. Finally, Chapter V concludes the paper.

II. RELATED WORK

A. CONSTRUCTION OF BSA

From the perspective of carriers’ network operations, they need to constantly maintain and update BSA. From the perspective of competition, they also need to know the BS deployment and coverage capability of its competitors’ network. Many third-party organizations also need to employ base station information to analyze mobile user behavior and provide various location-based service. Therefore, a complete and accurate BSA has great market demand, and as to how to obtain this information is also an important subject worthy of study.

We know that BSA describes the basic parameters of each cell in a network, and generally includes the BS name, cell name, site location (latitude and longitude), cell identity (CI), BS type, azimuth, downtilt, BS height, coverage scenario, etc. The encoding of CI is different for different networks. For example, in an LTE network, the cell is uniquely determined by a triplet \{TAC, eNBID, cellID\}, that is, tracking area code, base station ID, and cell ID. For 5G NR, it refers to \{TAC, gNBID, cellID\} instead. BS type refers to whether the base station to which the cell belongs is an omnidirectional one (usually only one cell) or a directional station (usually a base station contains multiple cells, also called sectors).

Traditionally, BSA data is manually measured and reported by the engineers. The disadvantages of this method are obvious. It is always human labor intensive, and may lead to manual errors during the process; further, large delay may be introduced during the reporting process.

In addition, BSA is not open for public. Therefore, the internet service providers such as Google and Apple have to collect the key BSA information through scanning from the terminal side.

B. SECTOR AZIMUTH DETECTION METHOD

Engineering parameters such as sector azimuth, downtilt, and site location are the key factors that affect the quality of network coverage. Therefore, carriers have put lots of manpower and resources to acquire and optimize them.

In addition to the manual on-site measurement of azimuth mentioned above, there are also some researchers who try to use other data sources for off-site azimuth estimation.

An azimuth estimation algorithm based on Gaussian distribution has been proposed in [1]. It utilizes the measurement report (MR) data collected in the base stations. Since the MR data sample usually does not contain the latitude and longitude information, it is necessary to use triangulation method to locate the samples first. Then, the distance between each sample and the BS, and the angle to north relative to the BS are calculated, and finally the azimuth of the sector is determined based on the Gaussian distribution statistics.

A linear regression algorithm has been utilized to estimate sector azimuth based on MR data in [2]. First, the distance between the sample and the BS is estimated according to the pathloss model. The sector coverage area is rasterized according to the sample distribution, and the average pilot signal strength of the Top N samples with the largest signal strength in each grid is calculated. Next, the samples with signal strength above the average are employed for the linear regression analysis to obtain the sector azimuth.

The assumption of the Gauss method is that the antenna deployment will be aligned with the users’ hot distribution area and thus, the sample density is the highest in the direction of the sector center. Actually, the sample distribution is largely affected by the geographical constraints (such as large buildings, roads and rivers, etc.), which may cause significant estimation errors. In addition, the above two methods are both based on MR data, and a major drawback of the MR data is the lack of precise positioning, which affects largely the estimation accuracy.

In [3], we proposed an azimuth estimation method based on radial rasterization (RR), which is an improvement of the Gauss method. But there is still much room for improvement in terms of the estimation accuracy.

C. MOBILE CROWDSENSING

Different from network-side measurement methods such as MR, with the rise of smartphones and wearable devices
equipped with various sensors, terminal-side measurement and applications based on crowdsourcing have received increasingly high attention of both the academics and industry [4], [5]. Ganti et al. [4] named this new paradigm of sensing as mobile crowdsensing (MCS). MCS is categorized into two types, namely, participatory perception and opportunistic perception. The former requires the active participation of individuals to contribute perceptual data, while opportunistic perception works in a passive and autonomous manner, and generally does not require active user participation.

The sensing data collected based on crowdsourcing can be applied to many fields, including network operation [6], traffic management, environmental monitoring [7], and daily life [8].

In this paper, we try to utilize the MCS data to study the azimuth detection based on machine learning.

The data was obtained through an MCS data acquisition platform. A data acquisition agent is deployed on a large number of smart phones to collect the LTE coverage information through the API provided by the operating system without interfering with the normal usage of the phone.

The sample mainly includes date and time of sampling, network type, TAC, eNBID, cellID, longitude, latitude, pilot strength (RSRP), and pilot signal quality (RSRQ). Generally, information such as phone number and text messages will not be collected, to avoid any violation of user privacy.

The longitude and latitude are acquired by two methods. In case GPS is enabled by the user, the precise GPS longitude and latitude are collected with a positioning error of several meters. Otherwise, the agent will acquire current location through third-party network augmented positioning APIs like Baidu or Google positioning. The network augmented positioning is less accurate (normally with errors of several tens of meters in urban area) and consumes less energy than GPS positioning. The proportion of the two methods in the MCS dataset is subject to the host APP and user behavior. According to the statistics of the MCS dataset employed in the following experiments, 73% of the samples are with GPS positioning and the remaining is with network augmented positioning.

D. MULTICLASS AND SOFT-MARGIN SVM

As the most well-known algorithm of statistical learning, SVM proposed by Vapnik exhibits outstanding performance for linearly separable binary classification [9]. For linearly non-separable problems, researchers have proposed soft-margin SVM [10], [11]. By introducing a slack variable to soften the interval, and penalty term C into the optimized objective function to punish the slack variables, a compromise can be achieved between the empirical risk and confidence risk.

A lot of work has been undertaken to further improve the soft-margin SVM. In Ref. [12], an additional slack variable called kernel slack variable for each quadratic constraint has been proposed, together with a novel soft margin framework for multiple kernel learning. Hajewski et al. [13] have proposed a new soft-margin SVM algorithm by utilizing a smoothing for the hinge-loss function, and an active set approach for the $\ell^1$ penalty. It enables to achieve a significant saving in computational complexity without sacrificing accuracy even for large dataset. In [14], it is reported that by incorporating data-based prior information into the black-box SVM model, the interpretability is enhanced.

Since many engineering scenarios like the one in this paper are multi-class classifications instead of binary ones, the standard SVM is extended to deal with multi-class issues. Generally, there are two types of methods, namely, direct and indirect methods.

The direct method is to modify the objective function directly by merging the parameter identification of multiple classification planes into one optimization problem, say, a piece-wise linear function, and solve the optimization problem “all together” [15]–[17]. It seems simple, but the computational complexity is relatively high, and thus difficult to implement. It is only suitable for small dataset and the accuracy is not significant either.

The indirect method is more popularly used in practice. It constructs a multi-class classifier by combining multiple binary classifiers, such as one-versus-rest (OVR), one-versus-one (OVO) [18] or hierarchical SVM [19]. OVR requires fewer classifiers (only M classifiers are needed for M-class task), and the speed is relatively fast. In [20], OVR is combined with decision tree to solve the multi-class problem. However, each classifier utilizes all the samples in the training, and the speed slows down sharply as the size of dataset increases; meanwhile, the negative samples are much less than the positive ones, and these causes an imbalance problem. The drawback of OVO method is that the number of binary classifiers increases as a quadratic function with respect to M, and the total training time and test time are relatively high.

As to which method is to be employed for multi-class problem depends on the scenario of application, such as the number of classes, data volume and the dimension of features.

III. AZIMUTH DETECTION ALGORITHM BASED ON SITE CONSTRAINT SOFT-MARGIN SVM

A. PROBLEM MODELING

In 4G and 5G mobile networks, the most common outdoor macro base station is sectorized BS with three sectors. Each sector covers roughly a range of $120^\circ$ by directional antenna, as shown in Fig. 1.

For a cell in a sectorized BS, its coverage area can be defined as a sector area centered on its azimuth. Since the antenna pattern is usually axisymmetric (see Fig. 1b), the azimuth is exactly on the axis of symmetry of the horizontal pattern. The problem of azimuth detection can then be transformed into finding the optimal boundaries of neighboring co-site cells.

In order to ensure the continuity of services when users move within the network, the boundaries of adjacent cells need to overlap to a certain degree to support the smooth
Each training sample corresponds to a slack variable, which represents the degree to which the sample does not satisfy the constraint, and \( C > 0 \) is a penalty term. The larger the value of \( C \), the higher the tolerance for the sample, and therefore, the smaller the decision boundary will be.

Another constraint that needs to be considered is, for each co-site sector, the boundary should go through the site location \( x_0 \). Then Eq. (3) can be reconstructed as:

\[
\min_{w, b, \xi} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \xi_i \\
\text{s.t. } y_i (w^T x_i + b) \geq 1 - \xi_i, \\
\xi_i \geq 0, \quad i = 1, 2, \ldots, N \\
w^T x_0 + b = 0
\] (4)

It is a convex quadratic programming problem, which can then be solved directly by common optimization calculation package. In practice, it is always converted into its dual problem by means of Lagrange multipliers instead.

Let’s introduce Lagrange multipliers for each constraint, i.e. \( \alpha = \{ \alpha_i \geq 0, \quad i = 1, 2, \ldots, N \} \), \( \beta = \{ \beta_i \geq 0, \quad i = 1, 2, \ldots, N \} \), and \( \gamma \geq 0 \). The Lagrange function of Eq. (4) is:

\[
\mathcal{L}(w, b, \xi, \alpha, \beta, \gamma) = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \xi_i \\
+ \sum_{i=1}^{N} \alpha_i \left[ 1 - y_i (w^T x_i + b) \right] \\
+ \sum_{i=1}^{N} \beta_i (-\xi_i) + \gamma (w^T x_0 + b)
\] (5)

Since the objective function with site constraint still fulfills the Karush–Kuhn–Tucker (KKT) conditions, it can then be transformed into its dual problem by Lagrange duality, that is,

\[
\max_{\alpha, \beta, \gamma} \min_{w, b, \xi} \mathcal{L}(w, b, \xi, \alpha, \beta, \gamma) \\
\text{s.t. } \alpha, \beta, \gamma \geq 0, \quad \alpha \geq 0, \quad i = 1, 2, \ldots, N \\
\beta_i \geq 0, \quad i = 1, 2, \ldots, N \\
\gamma \geq 0
\] (6)

Let us first find the minimum of \( \mathcal{L} \) with respect to parameters \( w, b, \) and \( \xi \), by partial derivatives.

\[
\frac{\partial \mathcal{L}}{\partial w} = 0 \Rightarrow w = \gamma x_0 + \sum_{i=1}^{N} \alpha_i y_i x_i
\] (7)

\[
\frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} \alpha_i y_i = \gamma
\] (8)

\[
\frac{\partial \mathcal{L}}{\partial \xi} = 0 \Rightarrow C - \alpha_i - \beta_i = 0
\] (9)

By substituting Eqs. (7)–(9) into Eq. (5), \( w \) and \( b \) can be removed, and the Lagrange function is rewritten as:

\[
\mathcal{L}(w, b, \xi, \alpha, \beta, \gamma) = \sum_{i} \alpha_i - \frac{1}{2} \sum_{i} \sum_{j} \alpha_i \alpha_j y_i x_i^T x_j
\]
\[
\begin{align*}
&+ \frac{3}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j + \frac{3}{2} \sum_i \alpha_i^2 x_i^T x_0 \\
\end{align*}
\]  
(10)

Subsequently, the objective function turns to be the maximization of the Lagrange function with respect to \( \alpha \), i.e.,

\[
\max_{\alpha} \mathcal{L} (w, b, \xi, \alpha, \beta, \gamma) \\
\text{s.t.} \quad \sum_i \alpha_i y_i = \gamma \\
C - \alpha_i - \beta_i = 0 \\
\alpha_i \geq 0, \quad i = 1, 2, \ldots, N \\
\beta_i \geq 0, \quad i = 1, 2, \ldots, N \\
\gamma \geq 0
\]  
(11)

By combining the conditions in Eq. (11), we have the final objective function as follows:

\[
\begin{align*}
&\min_{\alpha} \frac{1}{2} \sum_{i} \sum_{j} \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_i \alpha_i \\
&\quad - \frac{3}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_0 - \frac{3}{2} \sum_i \alpha_i^2 x_i^T x_0 \\
&\text{s.t.} \quad \sum_i \alpha_i y_i \geq 0 \\
&\quad 0 \leq \alpha_i \leq C
\end{align*}
\]  
(12)

By solving Eq. (12) we can have the optimal \( \{\alpha_i\} \) that minimizes the objective function.

Then the normal vector \( w \) can be obtained by substituting \( \{\alpha_i\} \) into Eq. (8) and (7). Suppose there are totally \( S \) support vectors among all the \( N \) samples, that is the samples \((x_s, y_s)\), \( s = 1, 2, \ldots, S \) which satisfy \( \alpha_s > 0 \), then the deviation \( b_s \) can be calculated by

\[
b_s = y_s - w^T x_s, \quad s = 1, 2, \ldots, S
\]  
(13)

Then we have the optimal \( b \) by averaging over all the \( b_s \), and finally the model of boundary classification \( f(x) = w^T x + b \) is achieved.

Since Eq. (12) is a quadratic programming (QP) problem, theoretically it can be solved by conventional QP algorithms, such as the Lemke method and interior point method. However, the computational complexity of solving Eq. (12) is proportional to the number of samples, which is very large in this case. Therefore, some highly efficient algorithms have been proposed specifically for SVM to find the optimal Lagrange multipliers, of which the most well-known one is the sequential minimal optimization (SMO) [21]. It takes the “divide and conquer” strategy. By transferring the highly sophisticated parallel optimization of multiple variables (i.e. \( \{\alpha_i\}, i = 1, 2, \ldots, N \)) into a series of dual-variable optimizations, a large portion of calculation is saved. Here, we employ SMO in solving Eq. (12).

Then, the question is “which method shall we take to deal with the multi-class problem for a multi-sector base station?”. Luckily for this case, knowing the left and right boundaries of a cell is sufficient to determine its azimuth.

Therefore, it is not necessary to find the boundary of every pair of cells in one base station, but those of adjacent cells only.

We call it the circular one-versus-one (C-OVO) method, as it is actually a variation of OVO. In this case, only \( M \) binary classifiers are needed for a \( M \)-sector base station, which is less than that needed for OVO in case \( M \geq 3 \). The number of classifiers is equivalent to that of the OVR method, but with no serious sample imbalance problem.

\section*{B. EVALUATION OF BOUNDARY SEPARABILITY}

To determine whether the two datasets (clusters) are absolutely linearly separable, we can construct the convex hulls of the two data sets and check whether they intersect. If not, the two are linearly separable. Generally, the Quickhull algorithm (http://www.qhull.org) can be used to find the convex hull of the data, and the Sweepline algorithm is employed to determine whether the edges of the convex hull intersect [22].

For linearly non-separable scenarios, however, although the soft-margin boundary can be found by allowing some samples to cross the boundary, the degree of confusion between the two sides still needs to be evaluated beforehand so as to determine whether it is worthwhile to classify. If the distribution of samples of the two clusters is highly confused, it makes no sense for classification. Otherwise, a soft-margin classification method can be used to estimate the optimal soft boundary. To this end, we define an index, namely, the average confusion index, to measure the degree of separability for adjacent co-site cells.

Suppose we have \( N \) cells (clusters) in a BS, i.e. \( C_1 \sim C_N \). Take the BS site as the center, we can divide the sample space into \( K \) subspaces by radial rasterization, and remove the subspaces whose sample size is less than a predefined threshold \( T_s \). The difference in the sample sizes of the two clusters in each subspace is the confusion index

\[
\text{con}(C_i, C_j) = \frac{1}{K} \sum_{k=1}^{K} \frac{|C_{ik} - |C_{jk}|}{|C_{ik}| + |C_{jk}|} \quad |C_{ik}| + |C_{jk}| > T_s
\]  
(14)

Here \( C_{ik} \) refers to the number of samples belonging to the \( i \)-th cell in the \( k \)-th subspace. Then the ACI is obtained by averaging the confusion index of all the cluster pairs, and represents the degree of separability of all the boundaries in a BS, that is:

\[
\text{ACI} = \frac{1}{N} \sum_{i,j=1, i \neq j}^{N} \text{con}(C_i, C_j)
\]  
(15)

The value of ACI ranges from 0 to 1. The larger the value of ACI, the higher the separability of the boundary.

It should be noted that, in case the sample sizes in the neighboring cells are highly imbalanced, the ACI calculated may be higher than it should be. To mitigate this influence, a random down-sampling shall be undertaken for the cells with much more samples than other co-site neighboring cells.
For the azimuth estimation task, for any BS with N sectors, the sectors are numbered in the clockwise order as C\(_1\sim C_N\). Then ACI can be calculated. If the index is greater than the preset threshold, the azimuth estimation is performed on this base station, otherwise it is ignored.

### C. AZIMUTH DETECTION BASED ON IMPROVED SOFT-MARGIN SVM

Based on the analysis given above, the detailed description of the azimuth detection algorithm is given below.

1) DATA PREPROCESSING

The input of the algorithm is the dataset crowdsourced in the live LTE network. It has advantages, such as larger data volume, broader temporal-spatial coverage, and more realistic reflection of the actual coverage of the base station. However, some data are poorly consistent, less accurate in positioning, and noisy. It may influence the accuracy of estimation. To this end, preprocessing on the original data is necessary.

First, the samples with missing or overflowed values are removed directly. Next, the isolated samples are filtered out by the Euclidean distance method. Finally, the longitude and latitude are normalized to facilitate the algorithm to converge faster, thereby improving the computational efficiency of the model fitting process.

Specifically, the spatial distance of each sample and the site location of serving BS are calculated. The outlier points which may be caused by positioning errors in the handset, or the points which are too far away from the site location will be identified and removed from the dataset by boxplot. The approximate distance calculation of Google Maps is used, i.e.,

\[
\text{dist} ((x_{11}, x_{12}), (x_{21}, x_{22})) = 2R \sin^{-1} \sqrt{\sin^2 \left( \frac{x_{12} - x_{22}}{2} \right) + \cos x_{12} \cos x_{22} \sin^2 \left( \frac{x_{11} - x_{21}}{2} \right)}
\]

where R≈6378137 is the radius of the earth in meters, and both longitude and latitude are in radian. Samples whose pilot strength are weak enough are also removed.

If the ACI of a base station is lower than the preset threshold, the base station is not processed.

2) ESTIMATION OF THE BOUNDARY OF ADJACENT SECTORS IN A BASESTATION

First, from the preprocessed dataset, all the samples with same eNBID are grouped together. We take the latitude and longitude as the attributes, and the cell ID as the label.

Suppose there are M cells in one BS (clock-wisely numbered as C\(_0\sim C_{M-1}\)), and the M linear boundaries obtained through SVM training of the C-OVO linear kernel with the method in Sec. III.A are Margin(C\(_i\), C\(_{mod(i+1,M)}\)), \(i = 0\sim M - 1\), that is:

\[
\text{Margin}(C_i, C_{mod(i+1,M)}) := w_i^T x + b_i = 0, \quad i = 0 \sim M - 1
\]

3) CALCULATION OF THE AZIMUTH OF THE SECTOR

After determining the boundary of adjacent sectors, the direction of the boundary vector needs to be further calculated. Specifically, it can be determined according to the sample distribution on both sides of the border perpendicular to the boundary and passing the site. Assume two neighboring cells C\(_1\) and C\(_2\) with N samples in total, the number of samples \(S_{up}\) and \(S_{down}\) on both sides of the border are:

\[
S_{up} = \{x_i \mid \frac{x_{21}}{x_{12}} > \frac{x_{22}}{x_{11}} (x_{i1} - x_{01}) + x_{02}, i = 1, 2 \ldots N\}
\]

\[
S_{down} = N - S_{up}
\]

The side with larger value is the direction of boundary vector.

Take the angle between the two boundary vectors of each cell as the span angle, then the vector equally dividing the span angle is the azimuth vector of the cell. The angle to north of the azimuth vector is the estimation of azimuth.

Take LTE network as an example, the pseudo code of the overall algorithm is shown in Fig. 3.

### IV. EXPERIMENTS AND ANALYSIS

In order to verify the performance of the proposed algorithm, a series of experiments are carried out by employing the network coverage dataset crowdsensed in the live LTE network. We then compare the performance with that of the Gauss and radial rasterization methods.

A. DATASET AND PREPROCESSING

Unlike the rich datasets for general machine learning study shared on the websites such as UCI, there is no MR or MCS dataset publicly available that can be used as the benchmark of algorithm validation in this paper.

Instead, with the support of local operator, the dataset employed in our experiments was acquired by the MCS data acquisition platform in the LTE network of Shanghai, China in 2017. For preprocessing and validation purpose, we have
also the BSA data which consists of the fundamental information of all the BSs in the network, including TAC, eNodeBID, cellID, site location, and azimuth of each cell.

The raw data acquired are first pre-processed. To begin with, the value of TAC, eNBID and cellID of each sample is searched in the BSA of the network. Those with invalid or null values are excluded from the MCS dataset.

Then we choose for validation only the samples of the tri-sector base station, which is the most popular type of outdoor base stations.

Next, all the samples of one BS are grouped together and the first and third quartiles are calculated. The samples whose distance to the site location is greater than the upper limit (i.e. $Q_3+1.5\times IQR$) of the boxplot is taken as outliers and removed from the dataset.

Take two base stations (eNBID = 372070 and 368718) as examples, the quartiles and upper limits of the samples are calculated, and their boxplot are illustrated in Fig. 4. The upper limits of them are 1061m and 1293m, respectively. Totally 1.6% and 1.9% of the samples whose distance to the site location are beyond the upper limit are removed.

Finally, after the normalization of longitude and latitude, we get the dataset for the experiments. It consists of 15,886,297 samples, covering 126,992 terminals and 1,250 tri-sectorized base stations.

The sample sizes of the cells are quite diversified, due to the free movement of users in the network. As seen in the CDF plot of sample size of all the 3750 cells (Fig. 5), around 50% of the cells are with a sample size of 3000 or more, and 30% of them have more than 5000 samples.

In the following experiments, the performance of the azimuth estimation is measured by comparing the estimation result with the actual value in the BSA. The metric is the difference between the two and it is based on a presumption that the azimuth in BSA is 100% correct.

First, the performance of the proposed method is validated for single BS, and compared with that of the other two methods. Then, we validate statistically the performance for multiple base stations. The impact of the key parameters and the sample size on the performance are also investigated.

### B. PERFORMANCE ANALYSIS OF SINGLE BS

First, two typical sectorized base stations are selected for the experiments, i.e., an urban base station E and a suburban base station F. The fundamental information regarding these is given in Table 1. Both of them are standard tri-sector BSs, with 54,092 and 29,404 samples, respectively.

Take the base station E as an example, after training by the site constraints soft-margin SVM algorithm with $C=0.8$, the hyperplane obtained and the corresponding maximum interval lines between neighboring sectors are shown in Fig. 6.
Further, the boundary vectors are determined using Eq. (18) and shown in Fig. 7. It can be seen that the boundary can be distinguished clearly for each sector.

Then we employ the Gauss, RR and SVM to estimate the sector azimuth of the base stations E and F. The results are shown in Fig. 8 and Fig. 9, respectively. The purple, green, and magenta points in the figure represent three sectors. The green arrowhead lines represent the true azimuths of each sector, and the dotted lines are the boundaries of neighboring sectors of the BS. The red, blue, and yellow arrowhead lines represent the estimated azimuth by the proposed SVM, the RR and the Gauss algorithm, respectively.

### TABLE 2. Error of estimation for the three algorithms (°).

| Base Station | cell 1 | SVM | RR | Gauss |
|--------------|-------|-----|----|-------|
| E            | 12    | 0   | 10 | 30    |
| F            | 5     | 10  | 15 | 30  |

Tab. 2 presents the error of estimation for each algorithm. For the base station E, the average errors of estimation are 5.74°, 4.67°, and 20°. For the base station F, the average errors are 6.7°, 13.3°, and 17.3°, respectively. It can be seen that the performance of the SVM algorithm is significantly superior to that of the Gauss and radial rasterization algorithms.

### C. IMPACT OF PENALTY TERM C ON PERFORMANCE

In Eq. (3), ξi is the classification loss of the i-th sample and it is 0 if the classification is correct. \( \sum_{i=1}^{N} \xi_i \) is the total loss. The smaller the loss value, the more accurate the classification of the training set. In principle, penalty term C can be any number greater than 0, as needed. A larger C indicates a higher requirement for precision in the entire optimization process, even at the expense of reducing the classification interval. When C approaches infinity, no classification error is allowed. It means a hard-margin SVM classification, which always leads to overfitting.

Here we take the base station E as an example to observe the error of estimation at different C settings. Fig. 10 illustrates the comparison of the estimated results for C values from 0.1 to 5. It can be seen that the performance is optimal at C = 0.8, but the performance difference is marginal.

### D. STATISTICAL ANALYSIS OF MULTIPLE BS

In order to have a more reliable observation of the performance of the three algorithms, we now consider the case when we have a large number of base stations.

1) IMPACT OF SAMPLE SIZE

First, we select 398 BSs out of all the 1250 BSs in our dataset, whose samples are above 5,000. The average estimation...
errors of SVM, radial rasterization, and Gauss algorithm are 25.47°, 31.07° and 32.31°, respectively. Therefore, the performance of the SVM algorithm is significantly superior to that of the Gauss and RR.

In order to compare and visualize the effect of sample size on each algorithm, the per-BS estimation error of these algorithms are illustrated in Fig. 11. It is seen that for all the three algorithms, the error of estimation goes down with increasing sample size in BS.

Now, we take a closer look at the impact of the sample size on the estimation performance (Fig. 12). Let us group the BSs into five categories according to their samples size; then, we see from Fig. 12 that the performance of the Gauss algorithm improves continuously with increasing sample size, while the RR shows only a slight improvement. Actually, it even gets worse for large sample size, which is probably due to the statistical bias. The SVM algorithm is in the middle.

By excluding the base stations whose sample size in any sector is less than 3,000, we select 124 tri-sector base stations. Here the ACI is not taken to exclude the base stations with severe sample confusion. The average estimation error of all the base stations for the SVM algorithm is 18.81°, which is significantly better than that of the other two (Gauss algorithm 28.5°, RR algorithm 23.5°). Fig. 13 illustrates the estimation error of each base station, where the base stations are ordered by the estimation error of SVM from left to right.

2) IMPACT OF BOUNDARY SEPARABILITY
It should be noted that due to severe spatial confusion in the samples of adjacent sectors, the error of estimation in some cases are as high as 40° or more, which is not acceptable in practice. The high confusion of samples and cell boundaries mainly comes from the very sophisticated scattering and reflection of wireless signals from BS. It always happens in dense urban areas with massive high-rise buildings.

One alternative solution is to exclude the BSs with highly confused cell boundaries according to their ACI. Below is an example of two BSs with very high or low ACI (Fig. 14). The ACI of the left one is 0.94, and the right one is 0.76. Obviously, the left one has more clear cell boundaries and achieves lower estimation error than the right one.

For different settings of ACI threshold, we recalculate the average and maximum estimation errors of the proposed algorithm, and these are shown in Fig. 15. For ACI≥0.9, the average error can be reduced to 16.6° and the maximum error kept within 35.07°. Therefore, it is suggested to set the threshold of ACI to be 0.85 or more in practice to guarantee a promising estimation performance.

Then what the percentage is for the base stations with an acceptable ACI? We calculate the ACI of all the 1,250 tri-sectorized BSs in the dataset, and the CDF of them is given in Fig. 16. It is seen that for 88% of the base stations, ACI is larger than 0.9, and 95.5% of them has an ACI of 0.85 or more.
E. ROBUSTNESS OF ESTIMATION

From the analysis given above, we can see that the proposed algorithm is quite sensitive to sample size, especially for the BS with samples less than 20,000. The larger the sample size, the lower the estimation error. In addition, the penalty term C has little impact on the estimation performance.

To further evaluate the robustness of the azimuth estimation, we calculate the standard deviation of estimation results for the 398 BSs. They are 19.2, 23.0 and 21.9 for SVM, RR and Gauss method, respectively. The proposed method is the most stable one among the three.

V. DISCUSSIONS AND CONCLUSIONS

As the key parameter of BSA, the azimuth of the sectorized base station is of great significance to carriers’ network operation. Traditional manual field measurement of azimuth has problems such as high cost, low efficiency and delay. With the popularization of smartphones, the mobile crowdsensing technology enables the acquisition of massive, real-time network information. With the help of data mining, a new way of sector azimuth estimation emerges.

Based on the analysis of the limitations of existing azimuth estimation methods, this paper has proposed a new method based on an improved soft-margin SVM algorithm. The explicit expression of the objective function of the problem is also deduced through Lagrange duality.

With the massive wireless coverage data crowdsensed in the mobile network, we have verified the proposed method and compared its performance with that of the two existing methods, namely, the radial rasterization and Gauss methods. The experimental results of single BS and multiple BSs show that the proposed method is significantly superior to the other two methods in terms of estimation accuracy. And the performance is significantly improved with the increase of sample size per BS, especially for the proposed SVM method and the Gauss method.

We have analyzed the impact of the penalty term C on the performance of the proposed method. Experiments have shown that the algorithm is not sensitive to the setting of C.

In addition, in order to evaluate the separability of sector boundaries, we defined a new indicator, namely ACI. By calculating ACI, we can effectively evaluate the quality of data and prejudge whether or not to estimate the azimuth for this BS. Experiments have shown that after removing the base stations with low ACI, the overall performance of the algorithm can be further improved, thereby improving the practicability of the algorithm.

It shall be noted that, only the MCS dataset is utilized in the experiments of this paper, due to the unavailability of MR data. Actually, both MCS and MR datasets are promising sources of data for azimuth estimation research. Each one has its advantages and disadvantages.

Generally, the MCS data has more precise positioning information than that of the MR data. The raw MR data has no positioning information, and the missing longitude and latitude are backfilled by post-processing algorithm [23]. The accuracy is subject to the positioning algorithm employed. A well-known one is the so-called triangle centroid location method whose positioning error is in the scale of 100∼200m. As mentioned before, the positioning precision of the MCS data is much higher.

The advantage of MR is the huge sample size than that of MCS. As we know, that the MR data is the measurement result of all users served by the MR-enabled base station. In theory, if the whole network is MR-enabled, the MR data of
all users in the network can be obtained. On the contrary, MCS is a sampling measurement, that is, only those users who have the host APP running on the phone can be measured. As shown in Figs. 11 and 12, with more samples the estimation accuracy will be improved.

Therefore, it is hard to say which kind of dataset is more suited for the azimuth estimation. It is quite complicated an issue concerning both technical and non-technical reasons such as availability and cost. It would be an interesting work to compare the azimuth estimation performance of the MCS and MR datasets, which is however beyond the scope of this paper.

The proposed method provides a new way of azimuth information acquisition for base stations in the whole network, and is a promising alternative to the conventional manual on-site measurement, with consequent large saving of cost and high efficiency.

In the future, we intend to further optimize the algorithm to improve the estimation accuracy. On the one hand, it is necessary to take full advantage of the other attributes in the data, especially the pilot strength and signal quality, to find the optimal boundary in a higher-dimensional space. In addition to the MCS data collected on the terminal side, other sources of data can also be used, especially the MR data acquired at the network side. How to combine these data with MCS data to better serve the azimuth estimation is one of the tasks that will be undertaken in the future.

REFERENCES
[1] S. Liang, F. Chen, and H. Song, “Optimal azimuth positioning algorithm for base station antenna based on Gaussian distribution,” Telecommun. Sci., vol. 34, no. 7, pp. 128–134, 2018.
[2] F. Gao, L. Lei, and R. Du, “Verification of the cell azimuth based on clustering of sampling point and division of switching quadrant,” Telecommun. Sci., vol. 33, no. 5, pp. 106–112, 2017.
[3] H. Wang, S. Xie, K. Li, and M. Ahmad, “Big data-driven cellular information detection and coverage identification,” Sensors, vol. 19, no. 4, p. 937, Feb. 2019.
[4] R. Ganti, F. Ye, and H. Lei, “Mobile crowdsensing: Current state and future challenges,” IEEE Commun. Mag., vol. 49, no. 11, pp. 32–39, Nov. 2011.
[5] L. Wang, D. Zhang, Y. Wang, C. Chen, X. Han, and A. M. Hamed, “Sparse mobile crowdsensing: Challenges and opportunities,” IEEE Commun. Mag., vol. 54, no. 7, pp. 161–167, Jul. 2016.
[6] P. Casas, M. Seufert, F. Wamsler, B. Gardlo, A. Sackl, and R. Schatz, “Next to you: Monitoring quality of experience in cellular networks from the end-devices,” IEEE Trans. Netw. Service Manage., vol. 13, no. 2, pp. 181–196, Jun. 2016.
[7] P. Dutta, P. Aoki, N. Kumar, A. Mainwaring, C. Myers, W. Willett, and A. Woodruff, “Common sense: Participatory urban sensing using a network of handheld air quality monitors,” in Proc. ACM SenSys, Nov. 2016, pp. 349–350.
[8] M. F. Butut, M. Demirbas, and H. Ferhatosmanoglu, “LineKing: Coffee shop wait-time monitoring using smartphones,” IEEE Trans. Mobile Comput., vol. 14, no. 10, pp. 2045–2058, Oct. 2015.
[9] V. Vapnik, The Nature of Statistical Learning Theory. New York, NY, USA: Springer, 2000.
[10] C. Cortes and V. Vapnik, “Soft margin classifier,” U.S. Patent 5 640 492, Jun. 17, 1997. [Online]. Available: http://www.freepatentsonline.com/5640492.pdf
[11] J. Shawe-Taylor and N. Cristianini, “On the generalization of soft margin algorithms,” IEEE Trans. Inf. Theory, vol. 48, no. 10, pp. 2721–2735, Oct. 2002.
[12] X. Xu, I. W. Tsang, and D. Xu, “Soft margin multiple kernel learning,” IEEE Trans. Neural Netw. Learn. Syst., vol. 24, no. 5, pp. 749–761, May 2013.
[13] J. Hajewski, S. Oliveira, and D. Stewart, “Smoothed hinge loss and ℓ 1 support vector machines,” in Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW), Singapore, Nov. 2018, pp. 1217–1223.
[14] S. Chen, C. Gao, and P. Zhang, “Enhancing interpretability of black-box soft-margin SVM by integrating data-based priors,” 2017, arXiv:1710.02924. [Online]. Available: http://arxiv.org/abs/1710.02924
[15] Y. Guemeur, “Combining discriminant models with new multi-class SVMs,” Pattern Anal. Appl., vol. 5, no. 2, pp. 168–179, Jun. 2002.
[16] K. Tatsumi, R. Kawachi, K. Hayashida, and T. Tanino, “Multiobjective multiclass soft-margin support vector machine maximizing pair-wise inter-class margins,” in Proc. Int. Conf. Adv. Neuro-Inf. Process. (ICONIP), Auckland, New Zealand, Nov. 2009, pp. 970–977.
[17] T. Tanino, R. Kawachi, M. Akao, and T. Tanino, “Performance evaluation of multiobjective multiclass support vector machines maximizing geometric margins,” Numer. Algebra, Control Optim., vol. 1, no. 1, pp. 151–169, 2011.
[18] U. Krelle, “Pairwise classification and support vector machines,” in Advances in Kernel Methods: Support Vector Learning. Cambridge, MA, USA: MIT Press, 1999, pp. 255–268.
[19] M. R. Azimi-Sadjadi and S. A. Zekavat, “Cloud classification using support vector machines,” in Proc. IEEE Geosci. Remote Sens. Symp. (IGASS), Honolulu, HI, USA, Jul. 2000, pp. 669–671.
[20] Z. Wang, C. Fang, F. He, Y. Zu, L. Wang, and J. Wang, “Hyperspectral rock spectral classification based on the decision tree-support vector machine (DT-SVMs),” Acta Scientiarum Naturalium Univ. Suntatseni, vol. 53, no. 6, pp. 93–97, Nov. 2014.
[21] J. Platt, “Sequential minimal optimization: A fast algorithm for training support vector machine,” Microsoft Res., Redmond, WA, USA, Tech. Rep. MSR-TR-98-14, 1998.
[22] S. Fortune, “A sweepline algorithm for Voronoi diagrams,” Algorithmica, vol. 2, nos. 1–4, pp. 153–174, Nov. 1987.
[23] S.-F. Chou, H.-L. Chao, and C.-L. Liu, “An efficient measurement report mechanism for long term evolution networks,” in Proc. IEEE Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC), Toronto, ON, Canada, Sep. 2011, pp. 197–201.
KE LI received the B.Sc. and M.Sc. degrees in applied geophysics from the Changchun University of Earth Science (now Jilin University), Jilin, China, in 1993 and 1996, respectively, and the Ph.D. degree in information and signal processing from Southeast University, Nanjing, China, in 2000, with honored award of provincial best Ph.D. thesis.

He has been with the Electrical and Computer Engineering Department, Concordia University, Montreal, QC, Canada, as Postdoctoral Research Fellow, from 2002 to 2003. He has been a Full Professor in telecommunication engineering with Beijing Union University, since 2013. Before that, he has been worked with Siemens China, JV of Huawei and Nokia Siemens Networks, and so on, for more than ten years, as a Senior Engineer, the Department Manager, and the Product Manager, engaged in the technical research, standardization, and product management of 3G and 4G mobile communication systems. He is the (co)author of more than 50 publications in academic journals and conferences and the holder of more than 30 national and international granted patents, in the area of signal processing, speech processing, mobile communications, and data mining. His current research interests include mobile big data, AI for network operations, network measurement, and cognitive radio.

M. N. S. SWAMY (Life Fellow, IEEE) received the B.Sc. degree (Hons.) in mathematics from Mysore University, India, in 1954, the Diploma degree in electrical communication engineering from the Indian Institute of Science, Bengaluru, in 1957, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Saskatchewan, Saskatoon, SK, Canada, in 1960 and 1963, respectively. He was conferred in 2009 the title of Honorary Professor at National Chiao Tong University, Taiwan. He is currently a Research Professor and holds the Concordia Chair (Tier I) in signal processing with the Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada, where he served as the Founding Chair of the Department of Electrical Engineering, from 1970 to 1977, and the Dean of Engineering and Computer Science, from 1977 to 1993. During that time, he developed the Faculty into a research-oriented one from what was primarily an undergraduate Faculty. He has also taught with the Electrical Engineering Department, Technical University of Nova Scotia, Halifax, and the University of Calgary, Calgary, as well as in the Department of Mathematics, University of Saskatchewan. He has published extensively in the areas of number theory, circuits, systems, and signal processing, and holds five patents. He is the coauthor of nine books and several book chapters. He was a Founding Member of Micronet from its inception, in 1990, as a Canadian Network of Centers of Excellence until its expiration, in 2004, and also its coordinator for Concordia University. He served as a member of the Board of Governors of the CAS Society. He is a Fellow of the Institute of Electrical and Electronics Engineers, the Institute of Electrical Engineers, U.K., the Engineering Institute of Canada, the Institution of Engineers (India), and the Institution of Electronic and Telecommunication Engineers (India). He was inducted in 2009 to the Provosts Circle of Distinction for career achievements. He was a recipient of many IEEE-CAS Society awards, including the Education Award in 2000, the Golden Jubilee Medal in 2000, and the 1986 Guillemin-Cauer Best Paper Award. Recently Circuits, Systems and Signal Processing (CSSP) has instituted a Best Paper Award in his name. He has served the IEEE in various capacities such as the President-Elect, in 2003; the President, in 2004; the Past-President, in 2005; the Vice President (Publications), in 1976 and from 2001 to 2002; the Editor-in-Chief of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I, from June 1999 to December 2001; an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, from June 1985 to May 1987; the Program Chair of the 1973 IEEE CAS Symposium; the General Chair of the 1984 IEEE CAS Symposium; and the Vice-Chair of the 1999 IEEE Circuits and Systems (CAS) Symposium. He has been the Editor-in-Chief of the journal CSSP since 1999.

***