Spectrum Occupancy Classification Using SVM-Radial Basis Function

Mitul Panchal(✉), D. K. Patel, and Sanjay Chaudhary

School of Engineering and Applied Science,
Ahmedabad University, Ahmedabad, Gujarat, India
mitul.panchal@iet.ahduni.edu.in,
{dhaval.patel,sanjay.chaudhary}@ahduni.edu.in

Abstract. With recent development in wireless communication, efficient spectrum utilization is major area of concern. Spectrum measurement studies conducted by wireless communication researchers reveals that the utilization of spectrum is relatively low. In this context, we analyzed big spectrum data for actual spectrum occupancy in spectrum band using different machine learning techniques. Both supervised [Naive Bayes classifier (NBC), K-NN, Decision Tree (DT), Support Vector Machine with Radial Basis Function (SVM-RBF)] and unsupervised algorithms [Neural Network] are applied to find the best classification algorithm for spectrum data. Obtained results shows that combination of SVM-RBF is the best classifier for spectrum database with highest classification accuracy appropriately for distinguishing the class vector in the busy and idle state. We made analysis-based on empirical SVM-RBF model to identify actual duty cycle on the particular band across four mid-size location at Ahmedabad Gujarat.

Keywords: Big data · Spectrum occupancy
Spectrum measurement · Communication · Machine learning
Classification

1 Introduction

The rapid growth of connected devices around the world has drastically increased the demand of wireless spectrum. Every wireless service needs a certain amount of spectrum for transmission of data. Although, Spectrum is a limited resource and expensive too, so we need to enhance current wireless spectrum capacity using the existing spectrum information. Monitoring explicitly, the current spectrum system has not been utilized perfectly. Therefore, it's very important to understand current trends of spectrum bands and to identify occupied or unoccupied slots with specific time interval on the spectrum band which would help to improve the current wireless system in a more advanced way and benefits to the opportunistic spectrum access policy. Big spectrum data is new resource for future wireless communication and it contains detailed information about the

P. Marques et al. (Eds.): CROWNCOM 2017, LNICST 228, pp. 112–127, 2018.
https://doi.org/10.1007/978-3-319-76207-4_10

spectrum behaviour and utilization. We use this big spectrum data for extracting some meaningful information using statistical methods which helps us to improve the spectrum sensing framework and recognize the requirement of spectrum band.

The term Big data describes the large volume of data both structured and unstructured [1]. Big data can be analyzed for insights that lead to a better decision and strategic business moves. Big spectrum data in [2], is a new resource for wireless communication to leverage the significance of spectrum and improve the spectrum allocation and enhance the capacity beyond the current scenario. Spectrum band holds two types of users, namely, the primary users (PUs) which have a licence for that band and other one is secondary users (SUs) which are non-licensed users. Various spectrum measurement campaigns are conducted for acquiring a wide range of the spectrum band. In [3–5] many spectrum measurements campaigns are performed to identify and understand occupancy statistics. Also in [6], the occupancy statistics were utilized in Singapore and identified those channel that has low or no active utilization. In [7], authors had covered frequency range between 804 MHz to 2750 MHz in urban Auckland, New Zealand. This analysis indicates that on average the actual spectral usage of the band is only about 6.2%. Furthermore, in [8], occupancy statistics were carried out in the band of 30 MHz to 3000 MHz in Dublin, Ireland. The results illustrates that the average spectrum usage during the measurement period was just 13.6%. In [9], an extensive measurement campaign conducted in Aachen Germany, compared indoor and outdoor results. They determined a very high spectrum occupancy in the indoor scenario in the band from 20 MHz up to 3 GHz. Considerably less occupancy was measured in the outdoor scenario. Similar work has also been conducted in [10], where measurement campaigns were made in the urban and rural area at Atlanta. In this measurement, they had done experiments in frequency range of 400 MHz to 7.2 GHz and spectrum occupancy was 6.5% where spectrum vacancy was 77.6% of the amount of white space in 5.4 GHz at urban area whereas 0.8% usage spectrum in rural area and 96.8% spectrum vacant in 6.6 GHz white space. In [11], the use of spectrum occupancy information is made to predict the channel status in the consequent time slot so that optimal spectrum sensing order can be achieved.

The proposed approach improves the throughput of the system while meeting quality of service. On other hand, machine learning technique is known for the most promising solutions in statistics. Machine learning algorithms are often heuristics, meaning they do not require prior or prerequisite assumptions of data. Hence, our objective is to investigate spectrum occupancy using ML algorithms and to acquire comprehensive knowledge about spectrum database. Foremost motivation is that machine learning has a different kind of strategies to find out the useful inference on data. The main contribution here is to formulate the model for prediction of spectrum usage class whether it is idle or busy channel using machine learning classification algorithms and prove that it's best fit method for spectrum database (Table 1).

Table 1. Measurement device details

Equipment	Specifications (MHz)	Remarks
RF Explorer	15–2700 and 4850–6100	Enhance capabilities
Nagoys telescopic NA-773	15–2700	HAM bands
Whip dipole antenna	2400–2500	High quality 2 dBi
Rubber duck antenna	5400–5900	Coverage 2.4 GHz band

The rest of paper is organized as follows: Sect. 2 focuses on measurement setup and methodology. Section 3, shows measurement scenario considered for this work. Spectrum occupancy characteristics are discussed in Sect. 4. Section 5, discusses different current methods for occupancy. The machine learning approach is discussed in Sect. 6. Results are compared in Sect. 6. Finally, Sect. 7 concludes the paper and discusses the future work.

2 Measurement Setup and Data

For our measurement setup, the equipment RF Explorer hand-held device in [12], is used for the spectrum acquisition. The range of RF Explorer is 15 to 6000 MHz. It's basically powerful hand-held hardware device which captures the spectrum data in real-time and also provides connectivity with a local laptop which has installed RF Explorer Windows software. This software provides a real-time visualization of data, trace export facilities and frequency monitoring. Figure 1 illustrates live GSM-900 data being captured on RF Explorer windows software.

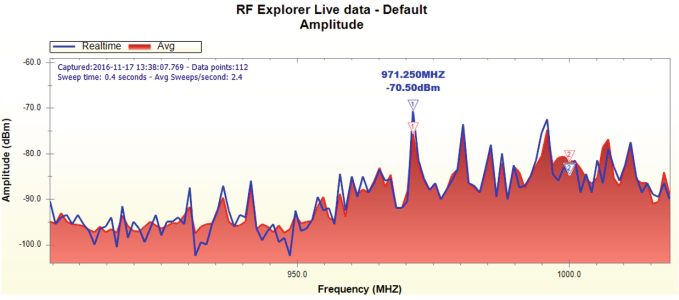

Fig. 1. RF explorer live receiving GSM-900 frequency

The measurement setup plays the key role for obtaining spectrum occupancy result because the percentage of accuracy result entirely depends on many levels of data accuracy. It's an essential part of every spectrum measurement setup. Here, measurement setup is carried out covering 820 to 960 MHz frequency which

lies in GSM-900 frequency band in ahmedabad, Gujarat. For analysis purpose, we have selected two GSM channel which lies in between 880–915 MHz (uplink) and 920–935 MHz (downlink), providing 124 RF channels with a bandwidth of 200 kHz. As mentioned in [13], there are some basic parameters that should clearly be specified like frequency name (GSM-900), location (four location of ahmedabad), direction (Omni-directional), polarization (Rx antenna is vertically polarized) and time variation (sampling rate 200 KHz and measurement duration 2 h).

3 Measurement Scenario

In this paper, we present the statistical analysis of spectrum occupancy in four most populated areas in Ahmedabad namely Nehrunagar, Navarangpura, Shivranjini and Vastrapur. Here considering only GSM-900 spectrum band and determining occupancy level in that band. The acquired data-set contains two main parameters one is power spectrum density (PSD) and another one is time variation (millisecond) which is continuously changing according to frequency power level in a spectrum band. The detail routes of spectrum data acquisition are shown in Fig. 2. At the other extreme, the minimum and maximum power level suggest the most frequent use of the signal. This difference would be separated by using thresholding algorithm. The approach of deriving spectrum occupancy is dependent on different parameters and methods; whereas the recent literature survey on spectrum occupancy measurement in [14] has proved that most frequent methodology for the spectrum occupancy measurement has Average Duty Cycle method, Markov Chain and Linear Regression. In the following section, we apply all the above methods for finding an occupancy statistics empirically with help of above data acquisition method (Fig. 3). The duty cycle can be calculated by following.

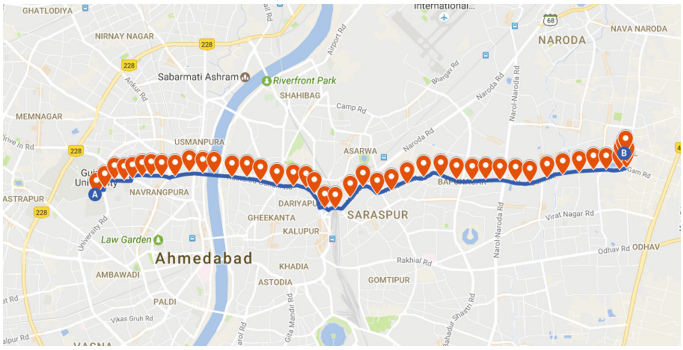

Fig. 2. A 44 location in Ahmedabad where spectrum measurement were collected

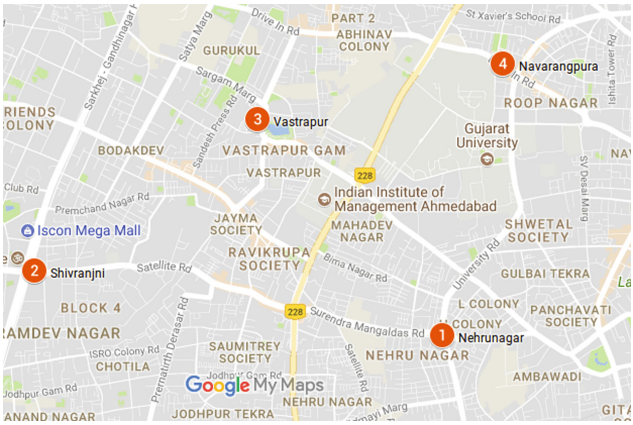

Fig. 3. A map of the four location in ahmedabad where spectrum measurement were collected

$$Duty\ \ Cycle = \frac{Signal\ \ Occupation\ \ period\,(n)}{Total\ \ observation\ \ period\,(m)} \times 100\%$$

where n represents time slot t, m denotes a total number of the time slot. Here received signal level is above the threshold.

4 Spectrum Occupancy Characteristics

In this model, all the parameters are used in terms of power level (dBm), because power spectrum density (PSD) corresponding to signal is different at all level or varies at all the times (Fig. 5). Thus, we have collected data from four areas Nehrunagar, Navarangpura, Vastrapur and Shivranjini. We measured common GSM-900 signal in all the locations and there are different utilization patterns acquired. Figure 4 represents the average power spectrum density in each location, moreover every location has different threshold value, but the range remains between −70 to −80 dBm. For threshold, we have taken average value

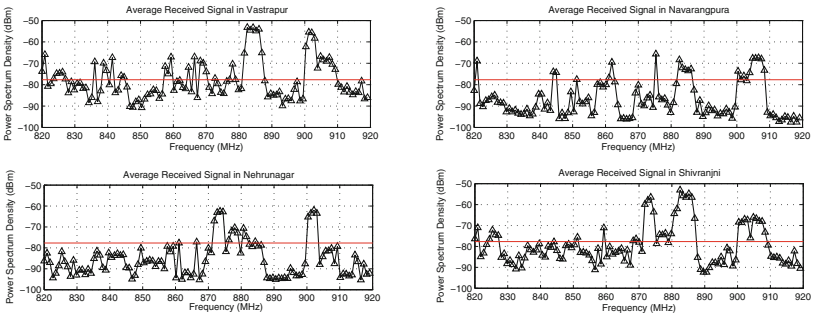

Fig. 4. Average power spectrum density at four location

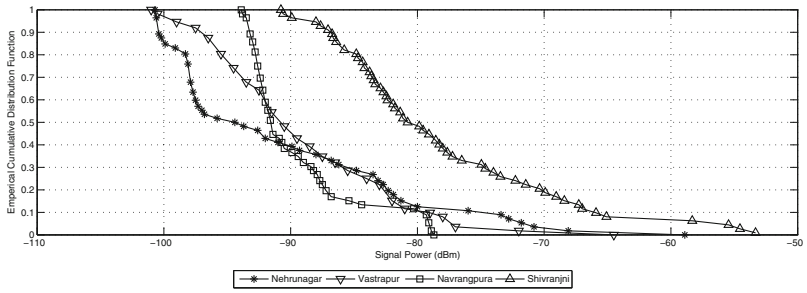

Fig. 5. Empirical cumulative distribution function showing spectrum occupancy

as threshold (λ) using Eq. 1, where $\tau_{(i)}$ is signal power and N is number of sample. In [7], these result helps us to identify the quantitative analysis. Figure 4, provides the overall trend of GSM signal occupancy irrespective of cities, time, sites and frequency and all the data points which were collected across the center frequency of 892.85 MHz. This cumulative distribution function shows the spectrum occupancy trends to each of the four location. It illustrates that Shivranjini is top most area where spectrum occupancy level is higher then Nehrunagar, Vastrapur and Navarangpura respectively. Using Empirical cumulative distribution it estimates the probability of each power spectrum density level in that location. Empirical CDF is a consistent estimate of the true CDF of any given value using Eq. 3. The empirical function $\hat{F}_n(t)$ gives weight of $\frac{1}{N}$ for each point of CDF, therefore it's also called step function. In below case, I is identical to X value, then weight of each value that is given to CDF is increased by $\mathrm{I}+1$. For every identical value of X, the given space of t is as follows

$$\lambda = \frac{1}{N}\sum_{i=1}^{N} \tau_{(i)} \tag{1}$$

$$I = \begin{cases} 0 & X_i \nexists t \\ 1 & X_i \neq X_{i+1} \Leftrightarrow Xi \leq t \\ I_{X_i \leq t} + I_{X_{i+1} \leq t} & X_i = X_{i+1} \Leftrightarrow X_i, X_{i+1} \leq t \end{cases} \tag{2}$$

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq t} \tag{3}$$

5 Methods for Spectrum Occupancy

In [15], different spectrum measurement campaigns have different goals varying from general analysis of spectrum utilization to specific individual wireless technologies. Furthermore, this section includes discussion on several important methodological aspects to be considered while evaluating spectrum occupancy.

Due to the differences in the signal modulation involved as well as the differences in the bandwidths utilized by each channel, energy spectral densities correspond to signals transmitted for different wireless services can be expected to be different.

This section gives the overview of all three methods which are used in this paper for deriving occupancy. Spectrum data is continuously generated through the device with a specific time interval. Linear correlation method finds the relationship between the parameter and finds the best fit variable which gives the relation between an independent variable and dependent variable [16]. This method gives a correlation metrics so using this we can model that and find the distribution of data.

Another method is duty cycle measurement which helps to find a specific load on particular frequency band at a particular time. The time interval is a most important role in duty cycle as it determines the average usage of frequency. The duty cycle methods are the most frequent methods for spectrum occupancy in [17]. Continuous Time Semi-Markov Chain (CTSMC) model, which is especially used to find patterns in real time database and classify the state in the database like idle or busy state. In [18], DTMC model is widely used in DSA/CR to describe the binary occupancy patterns in a channel in a time domain.

5.1 Linear Correlation Method

The normal linear model equation given in [19]:

$$\mathrm{y}_i = \beta_1 \mathrm{x}_{1i} + \beta_2 \mathrm{x}_{2i} + \ldots + \beta_n \mathrm{x}_n \tag{4}$$

$$+ b_1 \mathrm{z}_{1i} + b_2 \mathrm{z}_{2i} + \ldots + b_n \mathrm{z}_{ni} + \epsilon_i \tag{5}$$

$$\mathrm{y} = X\beta + \varepsilon \tag{6}$$

$$\mathrm{b}_i \sim N(0, \sigma D) \tag{7}$$

$$\epsilon_{ij} \sim N(0, \sigma \Lambda) \tag{8}$$

where

1. $y = [y_1, y_2 \ldots, y_n]^T$
2. X is the model matrix
3. $\beta = [\beta_1, \beta_2, \ldots, \beta_n]^T$ is the vector of regression coefficients
4. $\varepsilon = [\varepsilon_1, \varepsilon_2 \ldots, \varepsilon_n]^T$ is the vector of errors
5. N_n represents the n-variable multivariate normal distribution.
6. Λ and D are variance component

y_i is explains the relationship between one or more independent variables, called regressor variables, and a dependent variable, called the response variable (X). The parameters of the model are called the regression coefficients, specified as $\beta_1, \beta_2, \beta_3 \ldots, \beta_n$ and the error variance, defined as σ^2. The above model has one random-effect term, the error term ε_i given by

$$\varepsilon_i \sim N(0, \sigma^2)$$

5.2 Average Duty Cycle Method

Duty cycle [DC] model described in [18], where Ψ_t is able to describe its time evolution with sufficient level of accuracy. Notice that the DC in [20], is directly related to the instantaneous load or traffic load level supported by the channel. Although, traffic load would be experienced in a radio channel is frequently the result of an important number of random factors and aspects such as the number of incoming and outgoing users, the resource management policies employed in the system. The shape of Ψ_t, in this case, can be approximated by the summation of bell-shaped exponential terms centred at time instants t_m, with amplitudes A_m and widths σ_m given by:

$$\Psi_t \approx \Psi_{min} + \sum_{m=0}^{M=1} A_m e^{-(\frac{t-T_m}{\sigma_m})^2}, 0 \leq t \leq T \tag{9}$$

where $\Psi_{min} = min\{\Psi(t)\}$, T is the time interval over which $\Psi(t)$ is periodic (i.e. one day). The analysis of empirical data indicates that $\Psi(t)$ can accurately be described by means of M = 3 terms with τ_1 and τ_2 corresponding to busy hours and $\tau_0 = \tau_2 - T$. Notice that A determines the average value of $\Psi(t)$ in the time interval [0, T]. In [21], the occupancy level of various spectrum bands is quantified throughout this work in terms of the duty cycle.

5.3 Continuous Time Semi-Markov Chain Model

In [18], temporal spectrum occupancy pattern of a primary radio channel can adequately be modelled by means of a two-state Markov chain since it may be either busy or idle at a certain time instant. Let's denote $S = \{s_0, s_1\}$ the space state for a primary radio channel, with the s_0 state indicating that the channel is idle and the s_1 state indicating that the channel is busy. The behavior of a Markov chain can statistically be described with a set of transition probabilities among states. In the simulation of the Discrete Time Markov Chains (DTMC) channel occupancy model, the duration's of the sojourn times T were determined:

$$\mathrm{T}_i = \mathrm{N}_i * \mathrm{T}_s \tag{10}$$

$$\mathrm{P} = \begin{bmatrix} p_{00}\ p_{01} \\ p_{10}\ p_{11} \end{bmatrix} \tag{11}$$

The behavior of a markov chain can statistically be described as in [22] with a set of transition probabilities among states. If the state space S is finite with n states, the transition probability distribution can be described with a n × n square matrix.

$$P(t_k, t_l) = [p_{ij}(t_k, t_l)]_{n \times n} \tag{12}$$

To reproduce certain DC, indicated as Ψ in early, the transition probabilities of Eq. 11, that need to be satisfy some particular relations, determined as follows.

The n-element normalized row vector $\pi = [\pi_i]_{1\times n} = [\pi_0, \pi_1, \ldots, \pi_{n-1}]$, called the stationary distribution of the system, has elements representing the probability that the system is in each of its states in the long term, i.e.$\pi_i = P(S = s_i)$. For the occupancy model, where n = 2, the elements of π are given by [22].

Notice π_1 that represents the probability that the channel is in the busy state in the long term and it can thus be related to the channel's DC (i.e., $\Psi = \pi_1$).

$$\pi_0 = \frac{p_{10}}{p_{01} + p_{10}} \tag{13}$$

$$\pi_1 = \frac{p_{01}}{p_{01} + p_{10}} \tag{14}$$

There are many more approaches for spectrum occupancy model in [23] but here we applied Linear Mixed Effect Model on data. As mentioned earlier that, our data contains different site location, time, different power level, sweep frequency parameter which are captured by RF Explorer. One advantage of this model is, that it also works empirically with continuous data. A brief overview of Linear Mixed effect model is shown below.

5.4 Linear Mixed Effect Model for Spectrum Occupancy

The main advantage of linear mixed effect model are flexible, for instance enabling the modelling of altering slopes and intercepts. Linear regression has taken multiple input parameter which could be added in different dimensions e.g. space, frequency, time. For spectrum occupancy measurement, below Linear Mixed Effect Model is follows:

$$\begin{aligned} \text{Occ.Perc}_{ij} = \beta_0 + \beta_1 \text{Nav}_{ij} + \beta_2 \text{Neh}_{ij} + \beta_3 \text{Vast}_{ij} \\ + \beta_4 \text{Shiv}_{ij} + \epsilon_{ij} \end{aligned} \tag{15}$$

Linear mixed effect model takes all the parameter simultaneously. Consequently, all various location data matrix could be added in time and determine real-time spectrum occupancy. One significant benefit is that we can estimate spectrum occupancy in time domain as well as frequency domain and also build the prediction model. Here, we have calculated an occupancy rate for the different site. Table 2 shows the intercept of that model which is 0.72032 and applying linear mixed effect regression model (Eq. 15) for each location, we could determine occupancy rate in percentage. We could clearly see the spectrum occupancy in Navrangpura 0.72%, similarly for Nehrunagar 0.78% and so on. Table 3 illustrates result compression of spectrum occupancy rate using above three methods.

Figure 6 describes the residual plot versus fitted value. It is scattered plot of residual value on the y-axis and fitted value on the x-axis (estimated value). This plot is used to detect non-linearity, unequal error variances and outliers. The residual "bounces randomly" around 0 line. This suggests the assumption that the relationship is linear is reasonable. The residual roughly forms a "horizontal band" around 0 line. It also suggests that the variance of the error terms are equal. There have no one residual stands out from the basic random pattern

Table 2. Statistical parameter of linear mixed effect model

	Estimate	Std. error	tstate	DF	pValue
(Intercept)	0.72032	0.70323	1.0243	107	<0.05
Nehrunagar	−0.00046449	0.0028857	−0.16097	107	<0.05
Navrangpura	−0.00022675	0.0060755	−0.037322	107	<0.05
Vastrapur	0.0019761	0.0044519	0.44388	107	<0.05
University	0.0015222	0.003544	0.4295	107	<0.05

Table 3. Occupancy compression

Methods	Occ. rate	Std. error	Location
Liner mixed effect model	14.4%	0.7032	SEAS
Average Duty Cycle	40.3%	0.4801	SEAS
Discrete time Markov model	19.7%	0.6508	SEAS

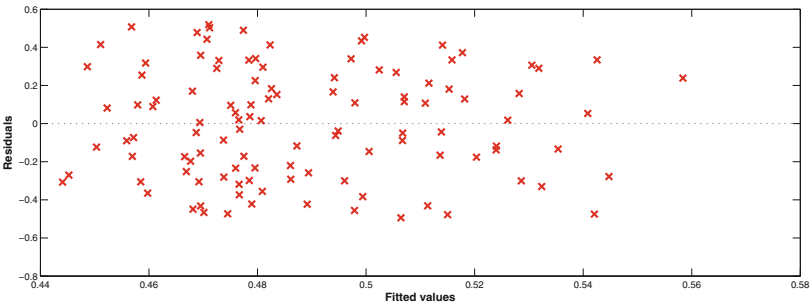

Fig. 6. Plot of residuals vs. fitted values

on residual, it means there are no outliers available. Here variance of residual increases with increasing fitted response which is knows as heteroscedasticity. This figure illustrates that there is some heteroscedasticity available.

$$Residual = Observed - Predicated$$

Consequently in Fig. 7, the corresponding is normal probability plot of the residual. This normal probability looks alike when residual is normally distributed and there is no outlier. It means this relationship is approximately linear with exception of few data point. We could proceed with the assumption that the error terms are normally distributed upon removing some outliers from data set and clearly see the points making diagonal line roughly straight. It means that 95% of residual is approximately fitted. Therefore, this data set follows the standard normal distribution.

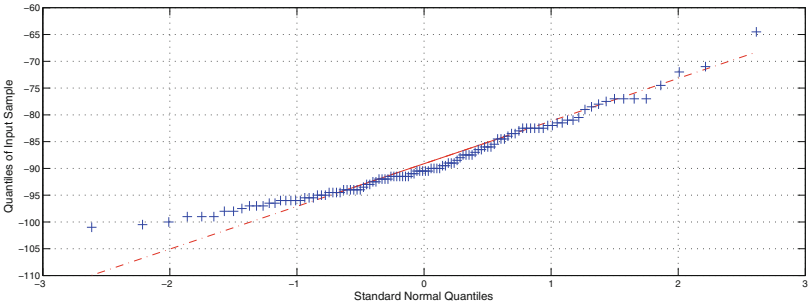

Fig. 7. Q-Q Plot of sample data versus standard normal

6 Statistical Analysis Using Radial Basis Function

Nowadays Machine Learning techniques are most important and fundamental things for big data analysis and statistics. Therefore, many researchers and academicians have acquired an interest in machine learning. There are two main parts in ML, supervised and unsupervised algorithms. These algorithms are most important for classification, clustering and prediction of future trends. Consequently, here we have use Machine learning algorithms to find an occupancy on spectrum data. In [11], machine learning techniques are used to identify occupancy in spectrum data and derive best-fit algorithms for spectrum data classification. Also, there are so many classification algorithms available such as Support Vector Machine, K-NN, Neural Network, Generalized Linear Model Decision Tree, Tree Bagger, Naive Bayes and so on. All these algorithms are used for classification purpose in a different areas. In this article proposed, SVM + FFT algorithm is used with which we obtained very good accuracy. But one major drawback is that it requires much computational time to train a large amount of spectrum data. So, moving a step ahead, we proposed SVM as it is advanced and requires less computational time. Also we explored the Support Vector Machine when using Radial Basis Function with scaling factor of 2, which derives classification result. We propose SVM-RBF model for calculating classification and its accuracy in MATLAB.

The work below depicts, how Redial Basis Function (RBF) with SVM combinations performs. For binary classification, given training data (x_i, y_i) for i = 1.....N, with $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$ learn classifier such that

$$f(x_i) \begin{cases} \geq \lambda & y_i = 1 \\ < \lambda & y_i = 0 \end{cases} \tag{16}$$

where λ value derived from Eq. 1 and $y_i f(x_i) \geq \lambda$ is busy state and $y_i f(x_i) < \lambda$ is idle state. The linear SVM classifier consists in defining function

$$f(x) = \sum_i (x_i^T, x) + b \tag{17}$$

which finds optimum hyper-plane and x^T weight vector which is normal to hyper-plane and b is bias value for weight vector. For non-linear, second solution is SVM with kernel function where kernel is a function that simulates the projection of initial data in a feature space with higher dimension $\Phi : K^n \rightarrow H$. In this new space data is considered linearly separable. By applying this, dot product and replacing value (x_i^T, x) in Eq. 18.

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right) \tag{18}$$

The new SVM-RBF function to classify the training data is:

$$f(x) = \sum_{i=0}^{N} y_i \ \alpha_i K(\mathbf{x}_i, \mathbf{x}') + b \tag{19}$$

where α_i Lagrange multiplier to solve problem easily.

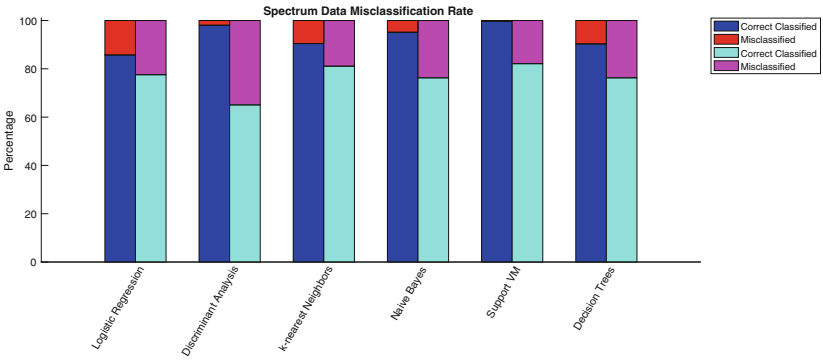

Fig. 8. Miss-classification rate of different classifier

Figure 8 illustrates classification accuracy of different algorithms. This figure represents training and testing set classification rate where violet and red colors indicate the training set correctly classified and miss-classified same as in testing set light pink and light blue indicates correctly classified and miss-classified rate in spectrum data. The updated SVM classifier (Eq. 19) algorithm use σ with value 2, where σ is scaling factor, whose value can change according to training set value. The confusion matrix describes the performance of different classification model on a set of test data for which the true value is known. The confusing matrix itself is relatively simple to understand, but the related terminology can be confusing. In order to understand, this figure explains different binary classifier's confusion. There are two possible predicated class, idle or busy. If it is predict idle it means signal is ideal, else signal is busy. The different binary classifier algorithms like Neural Network, Logistics regression K-NN, Decision Tree,

Naive Bayes and SVM-RBF are applied and test statistics is derived. In SVM-RBF, the classifier made a total of 972 predictions, for which classifier predicated 830 times idle and 142 times busy. In reality, they are 841 ideal and 131 busy samples. Overall, classifier correct classification rate is TP + TN/Total is 87.5%, so miss-classification rate is 12.5% and precision rate (correct prediction rate error) is 0.93%. This accuracy rate is very much higher compared to other classifier which indicates that SVM-RBF is the best classifier for Spectrum data. All the statistical accuracy-related result for algorithm is displayed in Table 4.

Table 4. Comparison of classification algorithm error

Classifier	Precision rate (predication rate)	Miss-classification error	Accuracy
Decision tree	0.85%	23.6%	76.4%
GLM	0.16%	23.4%	76.6%
Naive byes	0.42%	22.7%	77.3%
K-NN	0.84%	17.6%	87.5%
SVM-RBF	0.93%	12.5%	87.5%

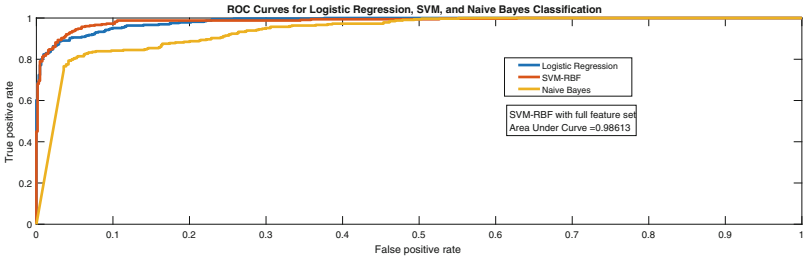

Fig. 9. ROC plot illustrate three discrete classifier

After classification of accuracy from the SVM-RBF we made receiving operating characteristics curve for cross verification, which is presented in Fig. 9. It's plot of true positive rate against false positive rate for different possible outcomes of a diagnostic test. It shows the trade-off between sensitivity and specificity. The closer the curve follows the left-hand border and top border of the ROC space, the more accurate test. The Fig. 9, illustrates three classifier algorithms performance. Using SVM-RBF function, the bounded line is to closer to 1 respectively logistic and naive Bayes classifier are decreased to lower bound. The area under the curve in SVM-RBF is 98.5% which means classification accuracy is very much higher than other algorithms.

Above spectrum analysis conducted with a supervised algorithm is based on a Support Vector Machine classifier. The algorithm analyses a set of positive

and negative class based on a frequency of occurrence in each class, estimating probabilities that value has positive and/or negative significance. Based on the probability of each value occurrence, channel or frequency probability is computed by calculating the product of that probabilities. This process requires pre-classification in two classes, Idle and busy data-set, specific to a supervised learning process, with which is calculated the occurrence of channel state in class. Obtained classification method also compared to simple SVM algorithm for the validity and efficiency of the SVM + RBF algorithm and it has outperformed to simple SVM algorithm. Table 5 illustrates the comparison result of the algorithms and it shows miss-classification rate and total classification loss are very low compared to SVM algorithms. In statistical classification, a confusion matrix is also known as an error matrix. It's specific table layout that generates a visualization of the performance of classification algorithm, typically in the supervised algorithm. Each column of the matrix represents the predicted class while each row represents the actual class. Confusion matrix has demonstrated in Table 6. To estimate the clustering results, accuracy, specificity, specificity, precision, recall, and F-measure were calculated over pairs of points. For a specific pair of points that share at least one cluster in the overlying clustering results, these measures try to estimate whether the prediction of this pair as being in the same cluster was correct with respect to the underlying true categories in the data. Precision is calculated as the fraction of pairs correctly put in the same cluster, recall is the fraction of actual pairs that were identified, and F-measure is the harmonic mean of precision and recall. Table 7 presented the statistical parameters for the efficiency of both the algorithms on spectrum data classification which is derived from the confusion matrix. In this study, for having an accurate assessment of the classification of two methods, it will be evaluated on the effectiveness, using same data set on which spectrum analytic is applied on frequency.

Table 5. Compression of SVM and SVM-RBF algorithms

Classifier	Sample size	Time (s)	Miss classification rate (MCR)	Loss
SVM	973	0.589	0.1182	13%
SVM-RBF	973	0.272	0.0946	7%

Table 6. Confusion matrix of SVM + RBF for efficiency

N = 973	**Predicated: Idle**	**Predicated: Busy**	
Actual Idle	TN = 238	FP = 75	313
Actual Busy	FN = 30	TP = 630	660
	268	705	

Table 7. Efficiency of algorithms

Algorithm	Accuracy	Sensitivity	Specificity	Precision	Recall	F-measure
SVM	0.8673	0.7692	0.9136	0.8081	0.7692	0.7882
SVM + RBF	0.8920	0.7596	0.9545	0.8876	0.7596	0.8187

7 Conclusion

This paper presents spectrum occupancy in GSM-900 band. We captured spectrum data of four different locations in Ahmedabad city. We analyzed traditional methods for spectrum occupancy and residual plot shown a good indicator of the occupancy but that methods are less accurate and required more computational time for the process of the spectrum data. Hence, machine learning comes in to the picture for the promising solution with less computational time and proposed a new method which derives the best accuracy an efficient way. By implementing SVM-RBF classifier algorithm, we determined that the SVM-RBF is the best fit for big spectrum data classification as it requires less computational time for training a data and demonstrates a good classification accuracy. This method could also be extended for different bands like ISM, Microwave, Satellites-Radar band and UHF-VHF band for opportunistic spectrum access.

Acknowledgment. We thank anonymous reviewers and our team members for the continuative support. This work was supported by Gujarat Council on Science and Technology, Department of Science & Technology, Government of Gujarat under the grant GUJCOST/MRP/2015-16/2659. The authors also thank Ahmedabad University for Infrastructure support.

References

1. Ding, G., Wu, Q., Wang, J., Yao, Y.-D.: Big spectrum data: the new resource for cognitive wireless networking. arXiv preprint arXiv:1404.6508 (2014)
2. Sasirekha, G., Dasari, S.R.: Big spectrum data analysis in DSA enabled LTE-A networks: a system architecture. In: 2016 IEEE 6th International Conference on Advanced Computing (IACC), pp. 655–660. IEEE (2016)
3. MacDonald, J.T.: A survey of spectrum utilization in Chicago. Illinois Institute of Technology, Technical report (2007)
4. Yucek, T., Arslan, H.: A survey of spectrum sensing algorithms for cognitive radio applications. IEEE Commun. Surv. Tutor. **11**(1), 116–130 (2009)
5. Patil, K., Prasad, R., Skouby, K.: A survey of worldwide spectrum occupancy measurement campaigns for cognitive radio. In: 2011 International Conference on Devices and Communications (ICDeCom), pp. 1–5. IEEE (2011)
6. Islam, M.H., Koh, C.L., Oh, S.W., Qing, X., Lai, Y.Y., Wang, C., Liang, Y.-C., Toh, B.E., Chin, F., Tan, G.L., et al.: Spectrum survey in Singapore: occupancy measurements and analyses. In: 2008 3rd International Conference on Cognitive Radio Oriented Wireless Networks and Communications, CrownCom 2008, pp. 1–7. IEEE (2008)

7. Chiang, R.I., Rowe, G.B., Sowerby, K.W.: A quantitative analysis of spectral occupancy measurements for cognitive radio. In: IEEE 65th Vehicular Technology Conference-VTC2007-Spring, pp. 3016–3020. IEEE (2007)
8. Tugba Erpek, K.S., Jones, D.: Spectrum occupancy measurements, shared spectrum company reports, shared spectrum company, January 2004–August 2005. http://www.sharedspectrum.com/wp-content/uploads/Ireland_Spectrum_Occupancy_Measurements_v2.pdf
9. Wellens, M., Wu, J., Mahonen, P.: Evaluation of spectrum occupancy in indoor and outdoor scenario in the context of cognitive radio. In: Cognitive Radio Oriented Wireless Networks and Communications, pp. 420–427. IEEE (2007)
10. Petrin, A., Steffes, P.G.: Analysis and comparison of spectrum measurements performed in urban and rural areas to determine the total amount of spectrum usage. In: Proceedings of the International Symposium on Advanced Radio Technologies (ISART 2005), pp. 9–12 (2005)
11. Azmat, F., Chen, Y., Stocks, N.: Analysis of spectrum occupancy using machine learning algorithms. IEEE Trans. Veh. Technol. **65**(9), 6853–6860 (2016)
12. Explorer, R.: Handheld hardware. http://www.wimo.com/rf-explorer-spectrum-analyser-signal-generator_e.html
13. Matheson, R.J.: Strategies for spectrum usage measurements. In: IEEE 1988 International Symposium on Electromagnetic Compatibility, Symposium Record, pp. 235–241. IEEE (1988)
14. Chen, Y., Oh, H.-S.: A survey of measurement-based spectrum occupancy modeling for cognitive radios. IEEE Commun. Surv. Tutor. **18**(1), 848–859 (2014)
15. López-Benítez, M., Casadevall, F.: Statistical prediction of spectrum occupancy perception in dynamic spectrum access networks. In: 2011 IEEE International Conference on Communications (ICC), pp. 1–6. IEEE (2011)
16. Pagadarai, S., Wyglinski, A.M.: Measuring and modeling spectrum occupancy: a massachusetts perspective. In: Proceedings of the International Symposium on Advanced Radio Technologies (2010)
17. López-Benítez, M., Casadevall, F.: Methodological aspects of spectrum occupancy evaluation in the context of cognitive radio. Eur. Trans. Telecommun. **21**(8), 680–693 (2010)
18. López-Benítez, M., Casadevall, F.: Discrete-time spectrum occupancy model based on Markov chain and duty cycle models. In: 2011 IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN), pp. 90–99. IEEE (2011)
19. Pagadarai, S., Wyglinski, A.: A linear mixed-effects model of wireless spectrum occupancy. EURASIP J. Wirel. Commun. Netw. **2010**(1), 1 (2010)
20. López-Benítez, M., Casadevall, F.: Spatial duty cycle model for cognitive radio. In: 21st Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 1631–1636. IEEE (2010)
21. López-Benítez, M., Casadevall, F.: Spectrum occupancy in realistic scenarios and duty cycle model for cognitive radio. Adv. Electron. Telecommun. Special Issue on Radio Commun. Ser. Recent Adv. Future Trends Wirel. Commun. **1**(1), 1–9 (2010)
22. Ibe, O.: Markov processes for stochastic modeling. Elsevier, Boston (2013)
23. López-Benítez, M., Casadevall, F.: Spectrum usage models for the analysis, design and simulation of cognitive radio networks. In: Venkataraman, H., Muntean, G.M. (eds.) Cognitive Radio and its Application for Next Generation Cellular and Wireless Networks. LNEE, vol. 116, pp. 27–73. Springer, Heidelberg (2012). https://doi.org/10.1007/978-94-007-1827-2_2