ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

ISSN 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

www.ijasem.org

Vol 18, Issue 1, 2024

A Comparative Study of Machine Learning Algorithms for Self-Localization Tasks

Mr. Adlakadi Anand

Assistant Professor, Department of CSE,

Malla Reddy College of Engineering for Women., Maisammaguda, Medchal., TS, India

Abstract—

The performance of a variety of machine learning approaches for use in localization systems is analyzed and compared in this study. An example of outdoor localization using multiple Received Signal Strength Indication (RSSI) values is examined, and the accuracy of outdoor localization is evaluated for a range of signal-to-noise ratios (SNR). The use of machine learning methods helps the system become terrain aware by automatically adjusting RSSI values in response to variations in the surrounding environment. In conclusion, this study shows a performance comparison of several classifiers that are accessible in the machine learning toolkit WEKA. The purpose of this comparison is to determine which of a set of models is the best suited for radio frequency propagation. According to the findings of our research, it is possible to correctly identify the terrain by using random forests and random committee classifiers, with an error limit of just 10 percent.

I.INTRODUCTION

One of the most important features of autonomous mobile devices is their ability to self-localize [1]. There have been many different approaches developed that can be used to integrate location awareness into systems [1, 12]. In this paper, an effort is made to enhance the location awareness of autonomous outdoor real-time locating systems that are based on the received Signal Strength (RSS) Indication [12]. Not only is the distance between transmitter and receiver a factor in RF propagation, but also the location of the receiver itself [4, 5, and 15].

The purpose of this study is to determine whether or not the machine learning tool known as WEKA (Waikato Environment for Knowledge Analysis) can be utilized to determine the classifier that has the highest possible chance of accurately predicting the correct propagation model. The capability of autonomous vehicles to recognize terrain will be improved if a self-localization technique and machine learning are combined into a single system. This will also result in an increase in the vehicles' overall localization effectiveness. Machine learning is an application of artificial intelligence (AI) that uses Statistical techniques to improve the performance of a system by providing it with data to automatically learn and develop from experience without being specifically programmed for that task. This can be accomplished without the system being specifically designed for that task.

According to our conceptualization of the system, this improvement would be brought about by the machine learning programmed WEKA (Waikato). Environment for Knowledge Analysis (also known as EKA for short) is a collection of Java-based machine learning software. We investigate the capability of WEKA's built-in classifiers to recognize a variety of RF propagation models and report our findings.

For the purpose of data mining, the following machine learning algorithms will be utilized: Bays Net, RBF Network, Star, Voting Features Interval (VFI), Decision Table Naive Bays (DTNB), Random Committee, and Random Tree.

To study RF propagation, different Hata models are implemented [6, 9]. The Hata model is an empirical model that provides the RSS variation depending on the distance travelled and the type of terrain.

The Hata Propagation model alludes to a variety of scenarios and mathematically describes what happens in each one. These are utilized in the process of predicting the propagation properties of a transmission wave based on a variety of parameters including the frequency of the wave, its distance travelled, the terrain, and other factors. The urban, suburban, and open-area scenarios are all being taken into consideration for this study. The variation in RSS that results from these circumstances is separated into training and test signals. On the basis of this signal strength, a test system is trained for terrain-learning. Following the analysis of the data, one can reach the conclusion that machine learning is responsible for the presence of artificial intelligence in autonomous localization systems. The findings are extremely encouraging, and they will make it possible to cut back not only on time but also on effort in the future.

II.METHODS

Classifier based on the A. Bayes Network

Bayesian networks are used most often in the categorization process. This is a probabilistic classifier that takes into account the training data in

order to determine the conditional probability of each attribute Ai given the class label C [19, 20]. By using the Bayes rule to determine the likelihood of event C, we can Given the specific occurrences of A1.....An, classification may be accomplished by first predicting the class with the greatest posterior probability, and then analyzing the results. When working with a set of predictors or traits, the objective of classification is to arrive at an accurate estimate of the value of a discrete class variable that has been assigned [21]. In specifically, the naive Bayes classifier is an example of a Bayesian network. In this particular model, the class does not have any parents, and each attribute has the class as its one and only parent [20, 21].

The Radial Basis Function, or "B,"

Radial basis function (RBF) networks use a static Gaussian function as the nonlinearity for the hidden layer processing components. These networks are also known as convolutional neural networks. Only a tiny portion of the input space around the point where the Gaussian is centered will elicit a response from the Gaussian function [22]. Finding appropriate centers for the Gaussian functions is the most important factor in ensuring the proper deployment of these networks [23, 24]. The training of an unsupervised layer is the first step of the simulation. The objective of this component is to generate the Gaussian centers and widths from the data that is provided as input. By using competitive learning [24], these centers are stored into the weights of the unsupervised layer. The widths of the Gaussians are estimated during the unsupervised learning phase depending on the centers of their neighbors. The output of this layer is obtained by applying a Gaussian mixture to the input data and then deriving the output from that.

C. Kstar

K* is an instance-based classifier, which means that it determines the category of a set of test data based on the training examples that are most similar to that set of test data, using some kind of similarity function. In contrast to other instance-based learners, it employs a distance function that is based on entropy.

D. Voting Options and Facilities Voting are used to determine the interval classification. Around each class and corresponding attribute, intervals are constructed. The number of classes assigned to each characteristic and interval are also recorded. The Decision Table and the Naive Bayes Model

A Decision Table and Naive Bayes hybrid classifier may be constructed and used with the help of the DTNB Class. At each stage of the search, the algorithm considers whether or not it would be beneficial to divide the characteristics into two distinct subsets: one for the decision table, and the other for the Naive Bayes method. It is a forward selection search that is carried out, and at each step, chosen attributes are modeled by Naive Bayes while the other attributes are represented by the decision table. In the beginning, however, all of the attributes are modeled by the decision table. At each stage, the algorithm takes into account the possibility of excluding an attribute from the model completely.

Random Committee, Form F.

It creates a collection of haphazardly chosen trees (base classifiers). The forecast is made by taking the average of the estimations of the probabilities. In order to guarantee the randomization of the basis classifiers, each base classifier uses a unique random number seed that is still derived from the same data [16]. Inducing randomness inside a system results in the generation of a heterogeneous ensemble of classifiers. The learner may be run several times with a variety of random number seeds, and then the predictions can be combined by averaging [10]. This approach is one technique to make the classification more reliable. The procedure is randomized by a random committee, which selects N of the best possibilities at random and then chooses the best option from that group.

G. Random Tree

The tree algorithms employ "divide-and-conquer" method. Nodes are produced at various stages of tree which test a certain characteristic. For numeric attributes, at each node a comparison of attribute is conducted with a preset constant which separates the tree into two additional branches [9]. This value is evaluated numerous times down the tree, every time comparing with a different constant. Randomness is produced by picking at random a collection of input characteristics to divide on [16].

H. the Hata Model is an empirical version of the route loss data supplied by Okumura [4]. It is used to anticipate the route loss over a wireless communication connection and is applicable for point-to-point and broadcast transmissions. The Hata Model offers a distinct model for varied situations. We are simulating data for WEKA analysis by employing the following three models. The data utilized in this inquiry is simulated in MATLAB. It contains a total of 3600 data points. 75% of the whole data is utilized for training and the remainder is used for testing. A quick introduction of the models utilized is provided below.

a) Hata model for urban areas

The Hata model for urban settings is the most extensively used transmission model for estimating cellular transmission in urban regions or cities where buildings and impediments in the line of transmission are stronger. This model accommodates for these

impediments by considering the impact of reflections, scattering and diffraction from these structures [7]. The coverage supplied by this model is for frequencies spanning from 150 MHz to 1500 MHz, mobile-station antenna height should be between 1 m and 10 m and base-station antenna height between 30 m and 200 m. Link distance reached with this model is between 1 kilometer and 20 km.

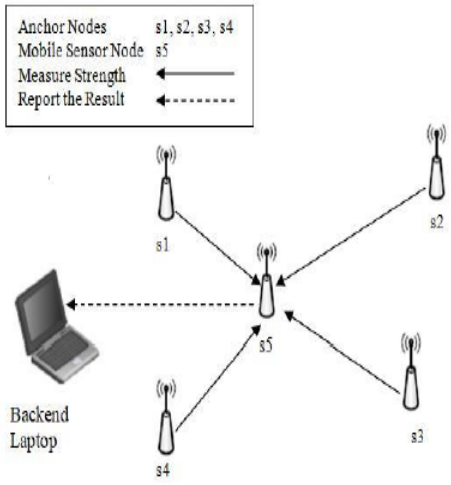
b) Hata model for sub-urban regions
The Hata model for suburban regions is an extension of Hata model for urban areas and predicts most accurate values for places that sit at outskirts of cities or where manmade objects do not present considerable transmission obstruction. It utilizes the metropolitan area propagation loss and transmission frequency to calculate the route loss.

III. WEKA

The University of Waikato in New Zealand is responsible for the development of WEKA, a data mining system that uses the programming language JAVA to carry out the implementation of data mining algorithms. WEKA is open-source software that is used to the process of creating machine learning (ML) methods and applying such approaches to data mining issues that occur in the real world. The methods are used by working directly with the dataset. WEKA is capable of implementing algorithms for data preprocessing, classification, regression, clustering, and association rules. The results of these algorithms may be examined immediately via the use of WEKA's visualization tools. With the help of this programmed, the user may also create their own own machine learning schemes [18]. The data file that serves as an input for Weak is saved in the ARFF file format, which includes specialized tags that identify various aspects of the data file (foremost: attribute names, attribute types, and attribute values and the data). The Explorer serves as the primary component of Weka's graphical user interface. It has a number of panels, each of which is capable of being used to carry out analysis on a dataset that has been loaded.

IV. SYSTEM MODEL

In a setting that is exposed to the elements, it is necessary to have an architecture that compiles data on the available signal strength. In a sensor network, the four reference nodes (s1, s2, s3, and s4) that are located in an exterior environment and have known placements are referred to as anchor nodes. This placement is shown in Figure 1. When a mobile sensing node (s5) is put into an environment, it immediately begins trying to locate itself by referring to the RSSI values (R1, R2, R3, R4) of the four anchor nodes while simultaneously broadcasting signals at a variety of RF frequencies. Both the anchor nodes and the sensor nodes all make use of the CC101 transceiver that we have. It is equipped with a register that can determine the RSSI value of the signal that is being received. These data are entered into a database after being read in over a TCP interface. An open space of 30 meters by 20 meters is used to collect the real-time data. The whole space is laid out in grids that are each 3 meters by 3 meters. The receiver device is monitoring the RSSI values of each anchor node at each grid point while the reference nodes are continually broadcasting signals at four distinct frequencies (433.1, 433.2, 433.3, and 433.4 MHz). The receiver node pinpoints its position by analyzing the intensity of the signals it receives from the other nodes in the network. The frequencies are referred to as "p1, p2, p3, and p4," and the identification of the terrain is determined based on the intensity of the signal. Modeling of the landscape is carried out since the signal strength is influenced by the topography in which the vehicle and reference nodes are located. When modeling the topography at these frequencies, the use of Hata models for open, urban, and sub-urban environments is recommended [7]. Using MATLAB®, we created a model that simulates the received power across a distance of 1.5 kilometers with a precision of 5 meters. The antenna height of the base station is decided to be 10 meters, while that of the mobile station will be 3 meters. The data set comprises 1200 values per model and a total of 3600 values. The Hata Model specifies the properties of radio frequency (RF) propagation, and this data collection is utilized for terrain learning based on those characteristics.



INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 18, Issue 1, 2024

Fig. 1. System Architecture

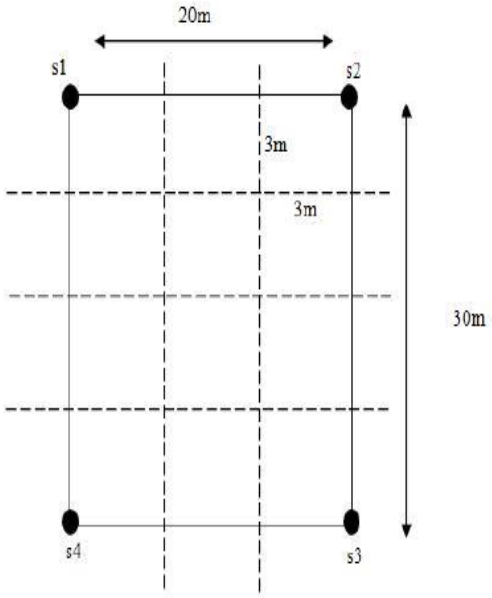


Fig. 2. An example of Node placement

V. SIMULATION RESULTS

In order to implement the appropriate HATA models, MATLAB® is used. The data set of the received power and frequency that corresponds to each model is written down into a text file that is then created. After that, this file is brought into WEKA so that categorization algorithms may be put into action. In order to study how well the various categorization methods work, we have split the data into two separate groups. The first batch of data, which contains 75% of the total original data, will be put to use in the training process, while the remaining data will be put to use in the testing process. Bayesian, functions, lazy, Meta, trees, rules, and other other classifiers may be used successfully on this dataset. The best results are achieved by utilizing meta-learner (Random Committee) and tree (Random Tree) classifiers while doing 10 fold cross-validations. The results of the simulation have been tabulated for ease of analysis, and they are only shown for the classifiers that provide the highest quality outcomes relative to their peers.

The classifiers are evaluated based on the number of instances that were correctly identified as well as the number of instances that were incorrectly identified expressed as a percentage. Following this, the mean absolute error, Root mean squared error, and Kappa statistics will only be expressed in numeric form. The term "error" is being used to refer to the disparity that exists between the real RSS value and the RSS value that was calculated by the WEKA classifier in question. On the dataset, we run a variety of algorithms, and then we tabulate the findings according to the number of properly detected instants. In order to classify the data, 10 fold cross-validations are performed. Among all of the other validation techniques, this validation strategy was shown to have the lowest bias and variance for the estimate accuracy [8]. As a result, it was selected as the method to use. The outcomes of the simulation are shown in the two tables that may be seen below. Table I is primarily concerned with providing a summary of the results depending on the accuracy of each simulation. Table II displays the results based on the errors that occurred while running the simulation. The graphical representations of the outcomes of the simulation are shown in Figures 3 and 4, respectively.

TABLE I SIMULATION RESULT OF EACH ALGORITHM

Classifier	Algorithm	Mean Absolute Error	Root Mean Square	Kappa Statistics
Bayesian Network	Bayes Net	0.1767	0.2977	0.685
Function	Radial Basis Function (RBF)	0.1921	0.3077	0.6563
Lazy	Kstar	0.209	0.3141	0.6567
Miscellanoeus	Voting Features Intervals (VFI)	0.2314	0.3146	0.6842
Rules	DTN B	0.177	0.2968	0.685
Meta-Learners	Random Tree	0.0412	0.2003	0.9083
Trees	Random Comitte	0.0412	0.2003	0.9083

TABLE II TRAINING AND SIMULATION ERRORS

Classifier	Algorithm	Correctly Classified Instants	Incorrectly Classified Instants
Bayesian Network	Bayes Net	79% (2844)	21% (756)
Function	RBF Network	77.08% (2775)	22.92% (825)
Lazy	Kstar	77.11% (2776)	22.88% (824)
Miscellanoeus	VFI	78.94% (2842)	21.06% (758)
Rules	DTNB	79% (2844)	21% (756)
Meta-Learners	Random Tree	93.88% (3380)	6.11% (220)
Trees	Random Committee	93.88% (3380)	6.11% (220

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 18, Issue 1, 2024

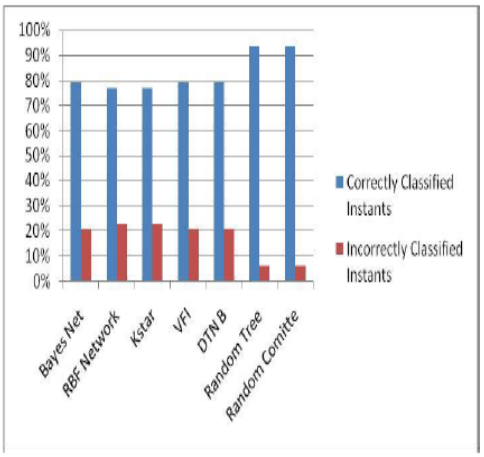


Fig. 3. Performance in terms of correctly classified instances

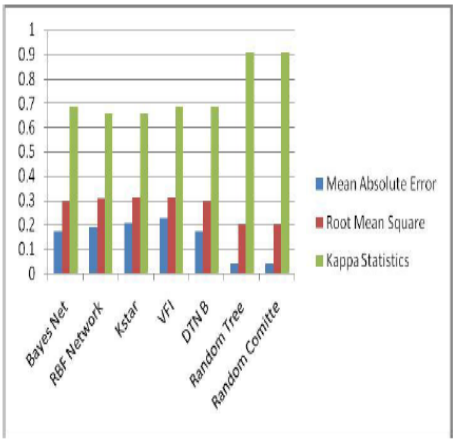


Fig. 4. Comparison between MAE, RMSE and Kappa statistics for classifiers used.

VI. DISSCUSSIONS

When we look at the data shown in the preceding Figures 1, 2, and Table 1, we observe that the maximum accuracy is 93.88%, and the lowest accuracy is 77.11%. The other algorithm achieves a level of accuracy that is around 78% on average. In point of fact, the Meta Learners and Tree classifier has the best accuracy, followed by the Bayesian Network and Rules Classifier with a percentage of 79%, and then the Miscellaneous with Lazy classifier. With a percentage that hovers around 77.08%, the Radial Basis Function classifier is at the very bottom of the chart. It was determined that out of a total of 3600 occurrences, an average of 2977 instances were properly identified, with the best score being 3380 instances and the lowest score being 2775 instances. In this simple experiment, which can be shown in Figure 2, we can clearly observe that When determining the accuracy of any given measurement scenario, it is customary to differentiate between the reliability of the data gathered and their validity [25]. The kappa statistic is used to do this assessment. The Kappa score obtained from the chosen method ranges between 0.6 and 0.7 on average. The accuracy of these categorization purposes is rather high [25], as determined by the standards provided by Kappa Statistic. From Figure 2, we can see that there are disparities in the mistakes that were caused by the training of the seven different algorithms that were used. This experiment suggests a widely used indication that is the mean of absolute errors as well as the root mean squared errors. In addition to this, the relative mistakes may also be used. Given that we have two different readings on the mistakes, it is prudent to choose the figure that is the average of the two. It has been determined that the RBF, Kstar, and VFI method has the greatest error rate. This algorithm has an average score of around 0.3, while the scores for the rest of the algorithms range from approximately 0.2 to 0.29 on averages. We will give preference to a method that has a lower error rate since this indicates that it has a more robust capacity for classification in terms of terrain recognition.

VII. CONCLUSIONS

In conclusion, we were successful in accomplishing our goal, which was to assess and explore seven different classification algorithms based on WEKA by making use of RSSI data values obtained by the Hate Model for urban, suburban, and open environments. The Random Committee and Random Tree Classifier algorithms, which have an accuracy of 93.88%, have been determined to be the best method based on the data produced by these models. When compared to the other classifiers, the Random Committee Classifier and the Random Tree Classifier have the lowest Average Error, which comes in at 0.2003. According to these findings, machine learning may be able to give our systems the appearance of intelligence by categorizing enormous data sets in accordance with a certain wireless propagation model. The Random Tree Classifier and the Random Committee Classifier are also options for determining the terrain's characteristics. In addition, since WEKA offers free and open-source classification techniques, our goal is to implement these classifiers in MATLAB® so that we can categories real-time data in a variety of environments. This may help us save a lot of time

and assist us in identifying the topography of newly collected data based on our past training and knowledge. In the future, one of our goals is to be able to create real-time values by taking measurements of the signal intensity at various locations within a distance of 1.5 kilometers and then classifying the results using WEKA classifiers. As a consequence of the real-world situation, the signal strength will be affected by the loss that is caused by external variables, which will result in a discrepancy between the simulated values and the real time values. Therefore, the accuracy of the classifiers may suffer when they are applied to data collected in real time.

VIII. REFERENCES

[1] K V. Zeimpekis, G. M. Giaglis, and G. Lekakos, "A taxonomy of indoor and outdoor positioning techniques for mobile location services," ACM SIGecom Exchanges, vol. 3, pp. 19-27, 2002.

[2] N. Bulusu, J. Heidemann, and D. Estrin, "GPS-less low-cost outdoor localization for very small devices," Personal Communications, IEEE, vol. 7, pp. 28-34, 2000.

[3] Bouet, M.; dos Santos, A.L., "RFID tags: Positioning principles and localization techniques," Wireless Days, 2008. WD '08. 1st IFIP , vol., no., pp.1,5, 24-27 Nov. 2008

[4] V. Abhayawardhana, I. Wassell, D. Crosby, M. Sellers, and M. Brown, "Comparison of empirical propagation path loss models for fixed wireless access systems," in Vehicular Technology Conference, 2005. VTC 2005-Spring. 2005 IEEE 61st, 2005, pp. 73-77.

[5] E. M. Van Eenennaam, "A Survey of Propagation Models used in Vehicular Ad hoc Network (VANET) Research," Pap. Writ. Course Mob. Radio Commun. Univ. Twenty, 2008.

[6] Y. Okumura, E. Homeric, T. Kawano, and K. Fukuda, "Field strength and its variability in VHF and UHF land-mobile radio service," Rev. Elec. Common. Lab, vol. 16, pp. 825-73, 1968.

[7] M. Hatay, "Empirical formula for propagation loss in land mobile radio services," Vehicular Technology, IEEE Transactions on, vol. 29, pp. 317-325, 1980.

[8] I. H. Witten, E. Frank, and M. A. Hall, Data mining: practical machine learning tools and techniques. Burlington, MA: Morgan Kaufmann, 2011.

[9] G. Biau, "Analysis of a random forests model," J. Mach. Learn. Res., vol. 98888, pp. 1063–1095, 2012.

[10] I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques: Morgan Kaufmann, 2005.

[11] M. M. Lira, R. R. de Aquino, A. A. Ferreira, M. A. Carvalho, O. N. Net, and G. S. Santos, "Combining multiple artificial neural networks using random committee to decide upon electrical disturbance classification," in Neural Networks, 2007. IJCNN 2007. International Joint Conference on, 2007, pp. 2863–2868.

[12] A. Dhurandhar and A. Dobra, "Probabilistic characterization of random decision trees," J. Mach. Learn. Res., vol. 9, pp. 2321– 2348, 2008.

[13] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," Syst. Man Cybern. IEEE Trans. On, vol. 21, no. 3, pp. 660–674, 1991.