2019 IEEE Pune Section International Conference (PuneCon)
MIT World Peace University, Pune, India. Dec 18-20, 2019

Towards Proxy-Attendance Detection using a Digital Attendance System and Machine Learning

Tanay Karve
Department of Computer Science
Fergusson College
Pune
tanaykarve@gmail.com

Rahul Dalal
Department of Computer Science
Fergusson College
Pune
rahuldalal99@gmail.com

Bharat Singh Dev Burman
Department of Computer Science
Fergusson College
Pune
bharat.singh4272@gmail.com

Smita Bhanap
Department of Computer Science
Fergusson College
Pune
srbhanap@gmail.com

***Abstract*—A digital solution to the attendance problem comprising of user interfaces, server technologies and a persistent database is presented. The system combines features of all the aforementioned technologies to maintain a record of students attending lectures scheduled in the respective university. The proposed system also incorporates an anomaly detection algorithm, backed by Machine Learning to ensure that the records entered accurately represent the students present in the lecture hall which therefore fortifies the system against unprofessionalism and malpractices. The system primarily consists of three main modules, namely student module, teacher module and server module. To explain further the anomaly detection process of the server module, plot graphs are provided which represent the position of students in the lecture hall documented by the system. This system overcomes the drawbacks of the canonical paper-driven attendance system, thus making the process swift and integral.**

Keywords—Machine Learning, Anomaly Detection, Attendance, PHP, MySQL, Sci-Kit Learn

I. INTRODUCTION

The proposed Digital Attendance System is designed peculiarly by considering the downsides and all the malpractices the traditional attendance marking system in most colleges are vulnerable to. Most colleges nationwide either stick to the policy of marking attendance by signing one's name in front of the respective names on a folio passed around by the professors generally after the emergence of lectures or the presence of every student is corroborated manually by the professor. Both the aforementioned methods are predisposed to unethical behaviour of students and are dilatory. To bridle these shortcomings the Digital Attend System is proposed which uses straightforward technologies which aren't much strenuous and are cost-effective. However, even though the underlying technology is simple, the problem of malpractices in the attendance process is quite complex. Many institutes are mandating fixed attendance criteria for students, which give rise to non-cooperation by the students and hence, malpractices. There is a rampant increase in attendance malpractices like signature forgery to log attendance of fellow students by their friends. This creates data inconsistency and violates many laws. However, normal models fail to validate attendance registration as there are new methods available every day to game the system. [1] There has been development of similar systems earlier by using biometrics or RFID technology. These systems were static and in a constant, unchanging environment. Due to a static system, the validations fail to learn the rapidly changing environment, which causes problems for both the students and the institution. Through the new proposed system, these factors are handled and a dynamic validator machine learning model is implemented, which continually learns the pattern of attendance registration based on GPS locations of the students. The algorithm we have implemented as a solution to this problem is the Elliptical Envelope model based on Gaussian distribution. The model fits a normal distribution to the attendance data and uses the Gaussian probability density function to predict the probability of anomalous registration as a function of GPS coordinates.

The system incorporates android applications which differ according to the role of the user (teacher or student), PHP which acts as a common gateway interface between the android app and the database to accomplish the task of logging in attendance records of the students and sending the synopsis of logged data back to the teacher, MySQL Database which stockpiles information regarding the users and the attendance register.

II. SYSTEM ARCHITECTURE

This system consists of three modules namely, the teacher client module, the student client module and the server module. Each module has a well-tailored control flow to regulate consistent functionality. [2] The motivation for building a mobile driven system was provided by similar systems built earlier and the rise in use and availability of Android smartphones in students.

A. The Student Module

This module consists of smartphones equipped with GPS and camera hardware. The QR codes, which are stickered on each of the benches, provide information about the lecture hall's unique ID and the bench's ID to the server. Each QR code can be scanned twice per attendance session, by two students. Location data, which is gathered from the

978-1-7281-1924-3/19/$31.00 ©2019 IEEE

smartphone, provides confirmation of the presence of the student. The location data along with the captured QR code data provides information to the server for marking attendance. QR codes have been earlier used in [3] for registration of attendance, but they represented details about lectures which was in session. Our use of QR technology is for obtaining bench and seating information from students.

B. *The Teacher Module*

This module consists of an internet connection capable device. No other hardware dependency is required and hence can be accessed from any device. This provides cross-user functionality along with device portability for ceaseless and incessant usage.

The system was built with user simplicity in mind. The clientele of our product consists of academicians and professors from a wide range of domains. They may or may not be tech savvy. Furthermore, this system provides a digital solution with swiftness as its core asset. The system's primary goal is to make the attendance process quick and integral, so that spending sizeable amounts of time for manual attendance wouldn't be entailed on the teachers. This mandates that execution speed and control flow must be robust, without forgoing on ease of use. Along these lines, the execution flow has been designed as follows.

To start an attendance session, the teacher selects the appropriate subject and the division or batch of the students. The lecture hall is implicitly selected according to the provided data by the teacher beforehand. It can be changed as per the requirement of the institution. The attendance session is started by making a request to the server using the application provided to the teacher.

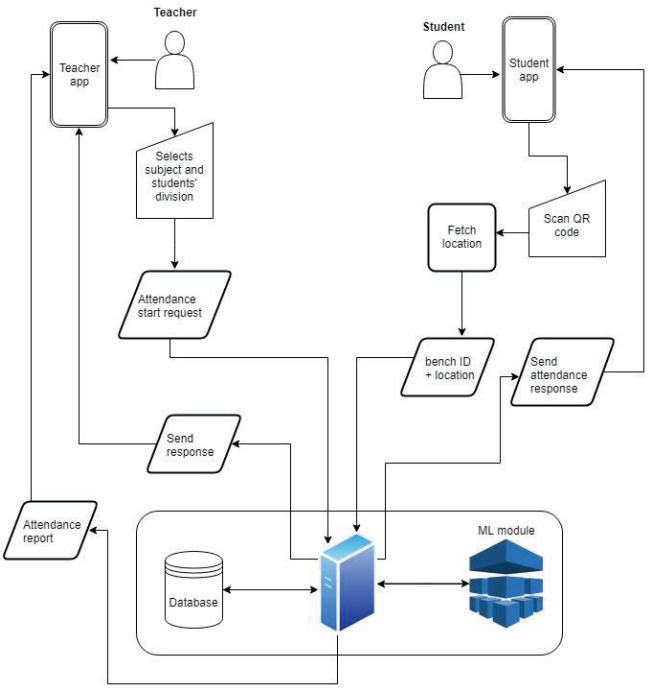
During this session, the students can give attendance requests to the server. Since sending requests doesn't require waiting for other students, this procedure can be done parallelly and this saves time and does not generate race conditions or data inconsistency. After an appropriate period of time, the attendance session can be requested to stop by the teacher, using the app.

The attendance session can be before the lecture starts or after the lecture ends, whichever is convenient for the teacher and students

C. *The Server Module*

This module consists of three submodules, namely the database module, the application programming interface (API) module and a machine learning (ML) module.

The database module, as the name consists of a persistent storage of attendance and users' data. All information, which is relevant and required for this system to work, is stored securely. The API module handles the sending and receiving of data to the clients, which are student and teacher modules. All data which is sent or received is done using a common network interface, which is JavaScript Object Notation (JSON). This allows for access calls to remotely stored procedures from clients independent of the client's platform.



III. CORE TECHNOLOGIES

The entire system is controlled by two primary governors, the QR code scanning functionality and the dynamic machine learning model. The QR reader, being employed along the front ends of the system, allows hands-free attendance registration to the students. The Machine Learning model, on the other hand does not face the end users and performs the procedural task at the system's back end. In order to detect anomalies in the registered attendance, this model is necessary as it provides a fluid platform which changes its validation according to the requirements of a dynamic, real-world environment.

A. *QR Parser*

A Unicode string or a binary data needs to be encoded in order to form a QR code. The following steps define the procedure in encoding a data string to a QR code [4].

Fig. 1. Sample QR code for Bench-ID

Algorithm:

Step 1: Choose an appropriate mode for encoding the text: numeric, alpha-numeric, byte and Kanji. Each mode uses different methods for conversion to string of binary bits.

Step 2: Encode the data using the appropriate coding mode for the text. The byte mode can encode any data but using alpha-numeric and numeric is more efficient if the data falls under these subsets.

Step 3: QR code includes error correction; some redundant data is created while encoding a QR code which helps the QR reader accurately read the code even if a part of it is unreadable. There are 4 levels of error correction which permits readability of code given that certain percentage of code is unreadable : L provides 7% error correction, M provides 15% error correction, Q provides 25% error correction and H provides 30% error correction. The scanner reads both the data codewords and the error correction codewords and determines whether it can read the data correctly or determine the errors if the data read is inaccurate.

Step 4: The data and error correction codes generated in step 3 are to be arranged in appropriate order. For larger QR codes these codes are generated in blocks which are interleaved according to the QR specification. This constitutes the structure of the final message.

Step 5: The binary bits are now placed in the QR code matrix. The finder patterns (boxes) are always placed on the top left, top right, and bottom left on the matrix. The separators (white modules) are placed next to the finder patterns. The alignment patterns(4 lines) and timing patterns (two lines) are also placed on the matrix depending on the specification. Before adding the data bits the dark modules are added which reserve area for format and version bits which are added later. Finally the data bits are placed into the matrix starting from the bottom right. The data bits incorporates white pixels for 0 and black pixels for 1.

Step 6: Certain patterns in the matrix make it difficult for the QR code reader to read the data and hence 8 masking patterns which alter the code according to their respective patterns. Each pattern has 4 penalty rules a particular pattern is chosen which has the least penalty score.

Step 7: The final step is to add version and format information by adding pixels to particular areas of the code that were previously left blank. The format pixels indicate the masking pattern and error correction level while the version pixels encodes the size of QR matrix used in larger QR codes

B. Machine Learning Model

It is the natural tendency of students to feel the urge to cut lectures when in school. However, the problem arises when a certain attendance mandate is fixed by the institution, requiring students to fill a necessary attendance limit. There are observable malpractices in the form of 'attendance-proxies' or just 'proxies' wherein, the students mark their attendance without actually being present for the classes. This is quite a conundrum for institutions as the regulation of attendance fails under this malpractice. A lot of measures have been previously taken, but most of them have certain fail cases. Our system provides an algorithm backed by Machine Learning to validate attendance for every single request. It uses appropriate GPS based feature engineering to learn and distinguish the true attendance locations from the false ones.

Similar problems are seen in server security, spam emails, engine diagnostics and others. These problems are solved by using anomaly detection algorithms which are able to determine if an event has occurred which is very different than the earlier events and thus an 'outlier'. The ML module is used to check for anomalies with the attendance data, which could be caused because of inconsistencies of the gathered attendance information and possible malpractices of 'proxies.' This module can improve over time with the increase in the amount of collected attendance data. The algorithm we have employed is the Elliptical Envelope anomaly detection algorithm.

The covariance estimation problem is a central part of multivariate analysis. It appears in many applications where we want to parameterize multivariate data. For example, in finance, the covariance matrix between assets' returns is used to model their risk. The covariance matrix is classically estimated using the maximum likelihood estimator (MLE) by assuming the data follows a multivariate normal distribution [6].

Let $R \in Rm \times n$ be a matrix of size m $\times$n that contains n observation vectors in n columns, each of length m. Under the normality assumption, the MLEs for the mean and the covariance matrix of these m attributes are:

Equation 1. Equation for MLEs for the mean and the covariance matrix of m attributes

$$\mu = \frac{1}{n} R * e.$$

$$\Sigma MLE = \frac{1}{n}(R - \mu e^t)(R - \mu e^t)^t$$

$$= \frac{1}{n}\Sigma (Ri - \mu)(Ri - \mu)^t \qquad (1)$$

where Ri is the column vector of size m for observation i and e is a column identity vector of size n of all ones.

It is assumed that the observations follow normality or the Gaussian function in the above estimation. It is thus inferred, that if even a single observation deviates from the gaussian distribution it would deteriorate the MLE estimators of the mean and the covariance matrix, which forms the basis for our anomaly detection solution. This algorithm thus generates the probability values from the probability distribution function or 'p.d.f' of the normal distribution as a measure of certainty, that the input supplied to the model, is not anomalous. The lesser the probability, the more confidence we have that there is anomalous behavior present. The 'p.d.f' of a normal distribution is given as follows.

Equation 2. Probability distribution function

$$f(x) = \frac{e^{-\frac{(x-\mu)^2}{(2\sigma^2)}}}{\sqrt{2\pi}} \qquad (2)$$

Where x is the input supplied,
μ is the mean,
σ is the variance

IV. ANOMALY DETECTION

This system along with the machine learning model was put into practice and employed at the Department of Computer Science, Fergusson College. The performance of the model was consistent and provided satisfactory results.

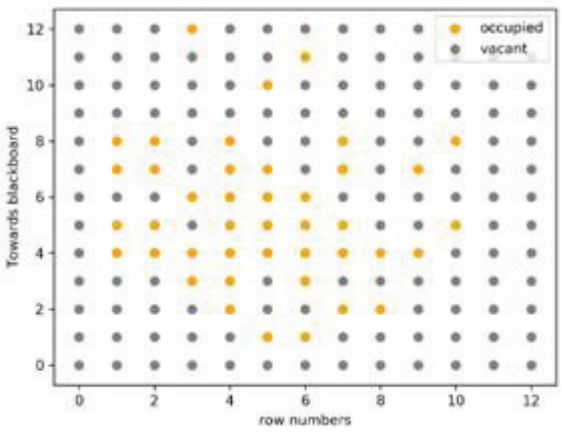


Fig. 2. Scatter plot from location data collected.

The above figure provides a concise idea about the seating problem. The orange markers belong to the location data collected from students as a training set for the model. The grey markers are vacant benches. For training, a dataset of eighty students was collected over a period of forty-eight lectures. It can be inferred that the seating arrangements follow a Poisson-like distribution. Consequently, our model can be fitted to this data, so that outliers can be detected. The GPS longitude and latitude coordinates are used as input vectors for the machine learning model. Hence, the model is able to determine if location coordinates are anomalous, due to students not actually being inside the classroom. The model parses each and every marked attendance entry for a particular lecture and alerts the teacher by sending a list of anomalous or possible 'proxy-attendance entries, done outside the classroom. The teacher can thus, call out the suspicious students, and if present, provide feedback to the model and if not, mark the student for malpractice. A demonstration of the model's working is as follows.

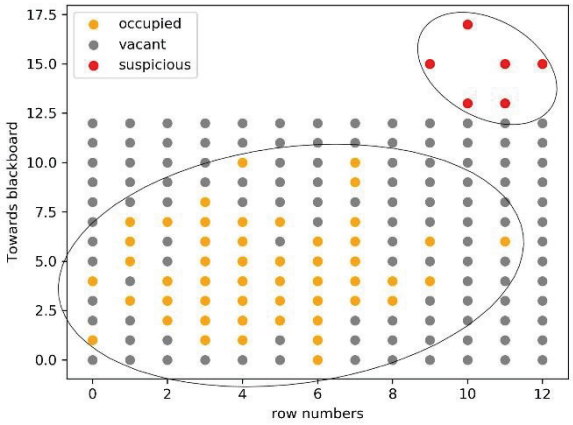


Fig. 3. Detection of anomalies by model.

Due to this ML model, the system can pick up on students who aren't inside the classroom and are marking attendance remotely and illegally. The main advantage of using an ML model over hardcoding location radius, is the automation features. It might save a lot of time for the institution to do other relevant work instead of manually determining a specific radius for each and every classroom.

A. Performance Metrics

1) Gaussian Distribution:

The probability of selecting scores from a given interval is also represented by the area under the curve above that interval [5]. After completion of preliminary data acquistion, a the data collected was observed to follow trends belonging to a Gaussian distribution. A Gaussian model was fitted to the data and the mean, variance and standard distribution was computed.

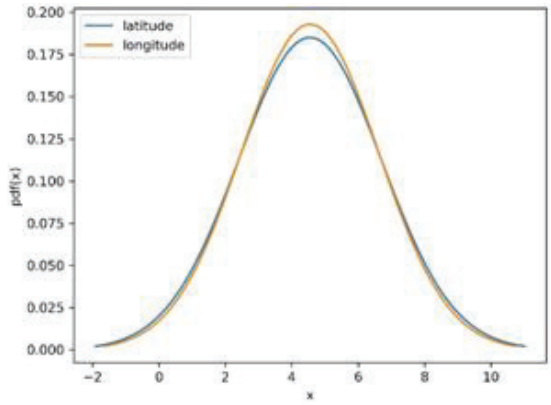


Fig. 4. Gaussian distribution for GPS coordinates.

TABLE I. NORMAL DISTRIBUTION PARAMETERS

Coordinates	Normal Distribution		
	Mean	*Variance*	*Std*
Latitude	4.55	4.6475	2.156
Longitude	4.35	4.2775	2.068

This table gives us a brief intuition, that the GPS coordinates of authentic users will lie around the latitudinal mean of 4.5 within 1 standard deviation of 2.15 and longitudinal mean of 4.35 within 1 standard deviation of 2.06. Consequently, it can be assumed that the 'proxy' or malignant users data lies outside these boundaries.

2) Anomaly Detection Model:

In machine learning and computational statistics, the accuracy of a statistical model is computed and interpreted by determining the values for sensitivity and specificity.

a) Sensitivity:

Sensitivity (also called the true positive rate, the recall, or probability of detection in some fields) measures the proportion of actual positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition) [7]. In our problem, True Positives (TP) was the total number of students' registration which was correctly identified by the system as normal and False Negatives were the total number of students' registration being anomalous but incorrectly classified as normal.

Equation 3. True Positive Ratio (TPR)

$$TPR = \frac{TP}{TP+FN} \tag{3}$$

$$= \frac{75}{75+1}$$

$$= 0.98$$

b) Specificity:

Specificity (also called the true negative rate) measures the proportion of actual negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition) [7]. In our problem, the False Positives (FP) were the students who were actually attending the class, but were detected as anomalous by the model and True Negatives (TN) were the students who were correctly detected by the system as anomalous.

Equation 3. True Negative Ratio (TNR)

$$TNR = \frac{FP}{FP + TN}$$

$$= \frac{2}{2+5}$$

$$= 0.28$$

V. Conclusion and Future Scope

This system was tested by simulating the environment of a designated classroom with eighty students over the period of forty-eight days. This allowed us to gain insights regarding applying our model system in the real life. The core advantage of our system is the automation of what was earlier thought to be a 'tedious' and unreliable process. The proposed system not only reduces paperwork for the institution, but it also takes care of attendance validation to provide true and persistent data as a part of the smart-classrooms initiative. Our future scope includes, but is not limited to real time deployment, ERP integration, seating arrangement analytics and seating revolution regulation i.e. maintaining rotation in seating arrangement, such that a particular student may not stick to one bench, thus maintain proper flow in the seating arrangements. We are sure that when deployed, this system will definitely overcome the problems of the canonical and age-old system of paper-based attendance.

Acknowledgment

We, the authors are grateful towards the help and support given by the Principal, Fergusson College (Autonomous) and the Department of Biotechnology for not only encouraging us to carry out implementation of the project, but also providing us with the funding required to develop and deploy the system under the DBT Star College Scheme. We would also like to thank our professor-guide Mrs. Smita Bhanap, for helping us throughout the process.

References

[1] T. S. Lim, S. C. Sim and M. M. Mansor, "RFID based attendance system," 2009 IEEE Symposium on Industrial Electronics & Applications, Kuala Lumpur, 2009, pp. 778-782.

[2] Hoda Abdelhafez, Maram Alamri, Riyof Alomari, Bayader Alzoman, Rfeef BinSheeha, Ayah Albawardi, Rehab, Mobile Based Attendance System Using QR Code, World of Computer Science and Information Technology Journal (WCSIT)

[3] Fadi Masalha, Nael Hirzallah, A Students Attendance System Using QR Code, (IJACSA) International Journal of Advanced Computer Science and Applications .

[4] S. Tiwari, "An Introduction to QR Code Technology," 2016 International Conference on Information Technology (ICIT), Bhubaneswar, 2016, pp. 39-44.

[5] Sue Gordon, The Normal Distribution, Mathematics Learning Centre University of Sydney NSW 2006

[6] Nguyen, TD. & Welsch, Outlier detection and robust covariance estimation using mathematical programming , R.E. Adv Data Anal Classif (2010) 4: 301. https://doi.org/10.1007/s11634-010-0070-7

[7] Wikipedia contributors. "Sensitivity and specificity." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 30 Aug. 2019. Web. 30 Aug. 2019