Dr. ShamimulQamar, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 12, Issue 11, November 2022, pp. 49-56

RESEARCH ARTICLE OPEN ACCESS

Naive Bayes Classifier based on pervasivecomputing for Classifying User Phone Call Behavior

Dr.Mohammed Inamur Rahman[1], Dr. Shamimul Qamar[2,2a], Dr. Abdulillah G F Saif[3], Dr. Magdi Mohammed Mohammed Ahmed Hamoda[4]
[1,2,3,4]*Computer Science & Engineering Department, College of Sciences & Arts, Dhahran Al Janoub Campus 64261 King Khalid University, Abha, Kingdom of Saudi Arabia, (KSA)*
[2a]*Meerut Institute Of Engineering & Technology, Meerut, India*

ABSTRACT:
The existence of noisy occurrences in mobile phone data is a major issue for classifying user phone call behavior due to the various potential harmful implications (i.e., accept, reject, missed, and outgoing). The classifier complexity may increase and the classification accuracy may decrease as a result of the amount of redundant training data. To To find these noisy cases in a training dataset, researchers use naive Bayes. By adopting the premise of independence and conditional probability of the attributes, the Bayes classifier (NBC) finds instances that have been erroneously classified. Meanwhile, some of these incorrectly classified events may be indicative of particular mobile phone users' usage patterns and behavioral inclinations. Existing Naive Bayes classifier-based noise detection techniques lack classification accuracy since this issue was not taken into account. In this article, we recommend a more advanced noise detection technique based on naive Effective Bayes classifier classification of user phone call behavior. By using both the discriminant Laplace estimator and classifier and the naive Bayes estimator, we dynamically calculate a noise threshold. We can identify loud circumstances using this noise threshold.
For the final decision-making process in this work, multiple Bayes classifiers are used and their user trustworthiness is discussed. The analysis of various common issues is followed by the provision of a pervasive computing architecture based on a basic but useful Bayes classifier model.
We employ the most popular classification algorithm (such as a decision tree) to evaluate how well our system classifies user phone call behavior (e.g., decision tree). Results from tests using a real phone The call log dataset serves as an example of how our suggested strategy, which more precisely detects the noisy instances from the training datasets, produces higher performance. In this study, we examine a scikit-learn implementation of the Gaussian Naive Bayes classifier, whose accuracy is 95%.

Keywords: Pervasive computing, Bayes classifier, ICT, Counting Trust, Counting Un-trust,Mobile Data Mining, Noise Analysis

---

Date of Submission: 01-11-2022 Date of Acceptance: 10-11-2022

I. Introduction

Our daily lives now frequently involve our mobile phones. The overall number of mobile cellular users is roughly the same as the world's population [13], and the owners of those phones spend most of the day with them as they go about their daily lives [13]. People utilize their mobile phones for a wide range of activities, such as voice communication, web browsing, app use, e-mail, online social networking, instant messaging, etc. [13]. Researchers have lately exploited a range of mobile phone data types for various specialized applications, including call logs [12], app usage logs [18], history of mobile phone notifications [11], browser logs [8], and context logs [23]. For instance, phone call data are used to predict user behavior to develop an automatic call rewall or call reminder system [14]. In the discipline of data mining, classification is a function that identifies and separates data classes or concepts [5]. Instances whose class values are unknown but whose attribute values are known are precisely identified using classification. It is possible to reliably determine user phone call behavior from log data using machine learning techniques like decision trees, but this is challenging since it requires a data set devoid of noise or outliers [3]. Noise, which is anything present in real-world datasets, might confuse the

relationship between an instance's features and its behavior class [6]. Such loud events could reduce classification accuracy and complicate the classification process. Additionally, it is evident that noise has a detrimental effect on decision trees [6].

The device we carry with us the most frequently is unquestionably a GSM-enabled cell phone. An antenna that covers a certain local area; an active connection, such as a call, to a particular antenna; and knowledge of the user's spatiotemporal position are all necessary for GSM communication between devices. The telecom company's collection of this data provides a spatiotemporal fingerprint of users travelling around a GSM coverage area. A user who only calls throughout the week and during business hours in a particular area, for example, can be considered a commuter as we only see his presence there during working hours. The observation made above prompts the following three queries: the first being that user habits and behavior are well-documented in GSM data; the second being that the volume of GSM data collected from the provider side is large, posing new problems for collection, storage, analysis, and mining; and the third being that the analysis of this enormous amount of personal data raises a number of privacy concerns.

This work suggests a way for understanding user behavior from call patterns by mining substantial amounts of GSM call data in order to address these three issues. A key element of our solution is the suggestion of a technique for behavior identification based on calling profiles of phone users. The pre-processed data in these call profiles can be used to run mining algorithms to identify various user activity categories.

The advantage of having call profiles created is that the analysis step is now based on an aggregated, privacy-preserving summary of the initial data instead of the original, massive, and sensitive GSM raw data. We show that these call profiles enable the development of a two-step process based on a bootstrap phase and a running phase for the classification of users into behavior groups based on their call behaviors.

More particularly, we provide a user behavior inference system that substitutes an improved inductive learning phase and automatic categorization for the two-phase learning approaches utilized in [1]. The call profiles are used as a quantitative model to estimate the amount of data provided from the TelCo operator to the data analyzer, which is another intriguing result of the analysis process's division into a bootstrap and operating phase. It is possible to construct a cooperation protocol based on the creation and analysis of explicitly defined pieces of information since both parties can utilize the call profiles as a common interchange model. Both sides gain from this in two important ways: first, the amount of data transferred is reduced by using pre-aggregated data, like call profiles, and second, privacy is better protected because no original raw data is given to the data analyst. Using call data provided by an Italian mobile phone company, the findings of a noteworthy experiment carried out in two Italian cities are discussed. Based on how people move around, we split users' call patterns in these studies into three groups: residents, commuters, and tourists. This might pave the way for a variety of cutting-edge uses, such mobility observatories, where monitoring systems regularly gather call patterns from the TelCo operator and then extrapolate customers' mobility behaviors.

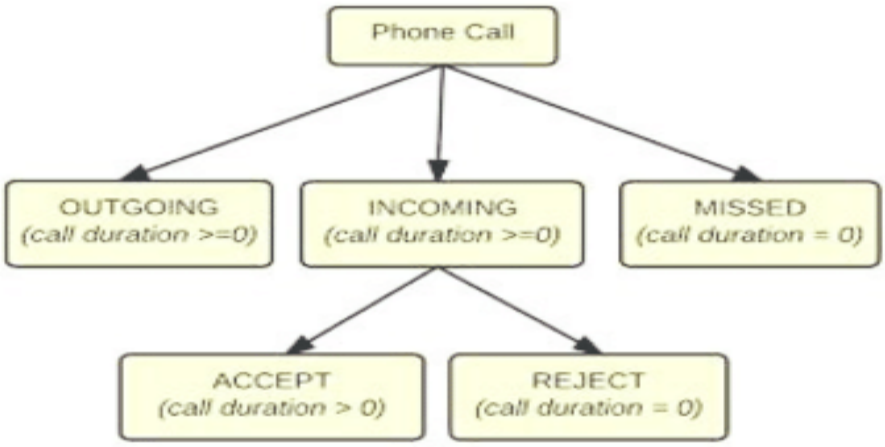

Figure1. Classifying User Phone Call Behavior

In order to categorize the user phone call behavior, we therefore characterize the effects of noisy cases as follows:
[1] The Increaseof the size of the rule-set by adding irrational classification rules that the users would not find interesting.
[2]The amount of training samples required and the complexity of the classifiers may both rise.
[3] Noise in the training data increases the likelihood that the decision tree classifier will overfit and lose accuracy.
The contributions are outlined as follows: We dynamically choose a noise threshold based on each person's particular behavioral tendencies.We provide an enhanced noise detection method based on naive Bayes classifier for accurately categorizing phone call activities of mobile users.
Our tests on actual mobile phone datasets demonstrate that this method is more effective than other methods for categorizing user phone call behavior.

II. Literature Survey

The historical backdrop of the research is provided in this section. This work expands previous findings by building on the earlier findings in a pervasive computing architecture based on a trust model that makes trust judgments dynamically depending on different contexts and sources of trust information[1,2]. Many initiatives in the field of trust modeling are built on the pillars of history, recommendation, and context[3]. Trust has been represented and calculated in a variety of ways based on the aforementioned techniques, including statistical analysis [4], probability [5], and directed graphs [6].

A simple probabilistic technique known as a naive Bayes classifier (NBC) can be used to predict the probability of class membership [2] [9]. Its ease of usage and the fact that probability generation only has to scan one set of training data are two of its main advantages. A naive Bayes classifier may easily tolerate missing attribute values by omitting the pertinent probability for those traits while calculating the likelihood of membership for each class. The conditional independence of classes is also demanded, which asserts that the effects of one attribute on one class are independent of the effects of other attributes.

In order to identify noise, we use the naive Bayes classifier (NBC) [9] as the fundamental tool. In order to ascertain the conditional probability for each attribute, we first use NBC to scan the training data. Table 1 displays a sample of the dataset for mobile phones. There are four attribute values and a phone call behavior for every instance (such as time, location, scenario, and caller-callee relationship). The prior probability for each behavior type and the conditional probabilities for each attribute value are shown for this dataset in Tables 2 and Table 3. Using these probabilities, we calculate the conditional probability for each case. similar to how NBC was established following independence.

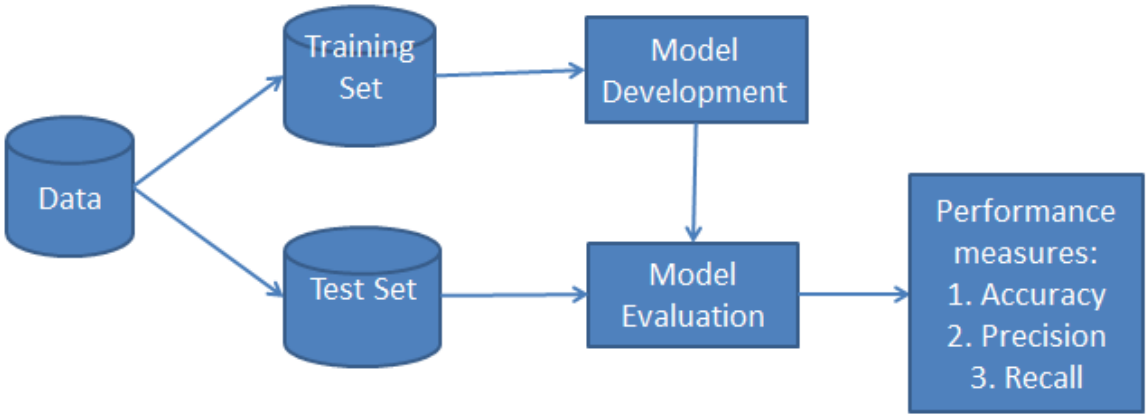

Figure2 : Classification Work flow of Naive Bayes Classifier

III. Naive Bayes Classifier

A classifier is a type of machine learning model used to distinguish between various objects based on specific properties. Ref[25].

3.1Principle of Naive Bayes Classifier:

A probabilistic machine learning model called a Naive Bayes classifier is utilized for classification tasks. The Bayes theorem serves as the foundation of the classifier.

3.2Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{--------(1)}$$

When B has already happened, we may use the Bayes theorem to calculate the likelihood that A will also occur. Here, A is the hypothesis and B is the supporting evidence. Here, it is assumed that the predictors and features are independent. That is, the presence of one feature does not change the behavior of another. The term "naive" is a result.
Bayes theorem can be rewritten as:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \quad \text{----(2)}$$

The class variable (play golf) in the variable y indicates whether the conditions are acceptable for playing golf or not. The parameters/features are represented by variable X.
X is given as,

$$X = (x_1, x_2, x_3, ..., x_n) \quad \text{---(3)}$$

Here, the features are represented by the numbers x 1, x 2,.. x n, which can be translated to outlook, temperature, humidity, and windy. By replacing X and expanding using the chain rule, we obtain,

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)} \quad \text{----(4)}$$

Now you may look at the dataset to get the values for each and then enter them into the equation. The denominator does not change for any of the entries in the dataset; it remains constant. As a result, the denominator may be eliminated and proportionality may be added.

$$P(y|x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y) \quad (5)$$

3.3 Types of Naive Bayes Classifier:
Multinomial Naive Bayes:
This is mostly used for document classification issues, such as determining whether a document falls under the sports, politics, technology, etc. category. The frequency of the terms included in the document is one of the features/predictors that the classifier uses.

Bernoulli Naive Bayes:
Similar to the multinomial naive bayes, but using boolean variables as predictors. The only options for the factors we use to predict the class variable are yes or no, as in whether a word is in the text or not.
Gaussian Naive Bayes:
We assume that the values of the predictors are samples from a gaussian distribution when they take up a continuous value rather than being discrete.

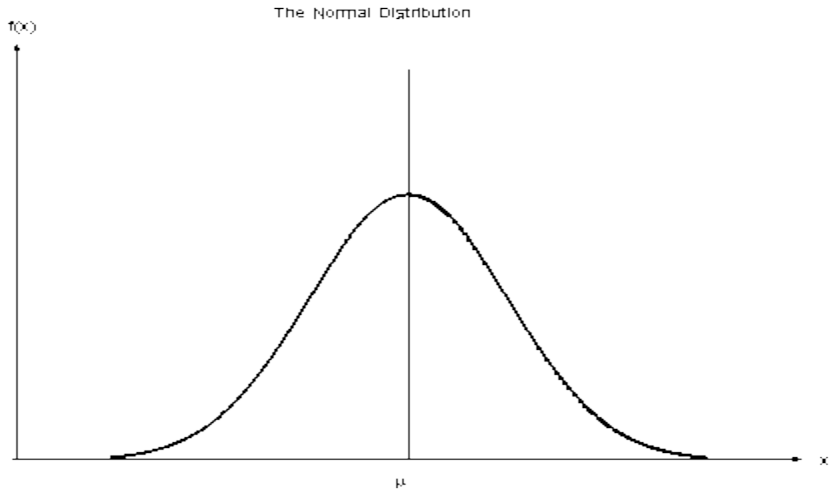

Figure 3. Gaussian Naive Bayes

The conditional probability formula changes to, since the dataset's presentation of the values changes.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad 6)$$

IV. MATERIALS AND METHODS

The research's materials and methodology are presented in this part.

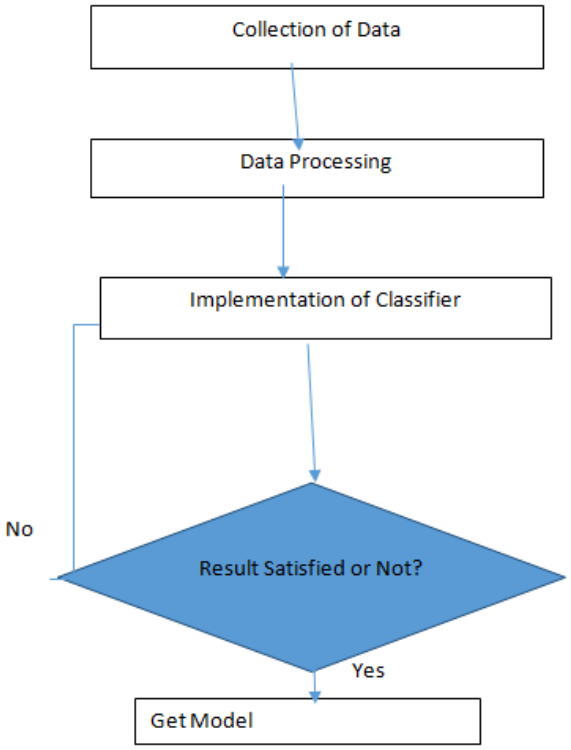

Figure 4. Proposed Method

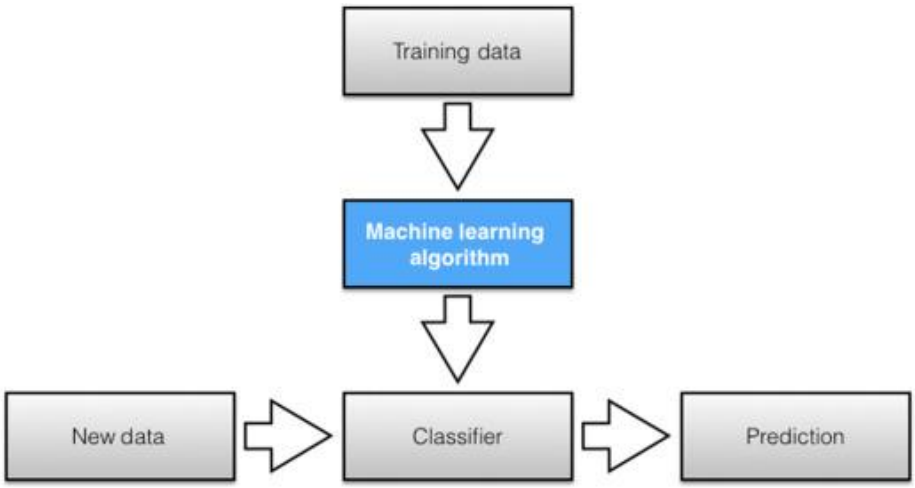

Figure5: Classifier data Training Using Machine Learning

V. Data Set Description:

The theory underlying the Naive Bayes classifiers and how they work are covered in this article.

A group of classification algorithms built on the Bayes' Theorem are known as naive Bayes classifiers. It is a family of algorithms rather than a single method, and they are all based on the idea that every pair of features being classified is independent of the other.

The data was developed based on three different sorts of attacks. Based on counting-based, time-based, and context-based attacks, the user develops a reputation. The suggested method makes it possible to assess each user's reliability by keeping an eye on how they interact with one another on the network

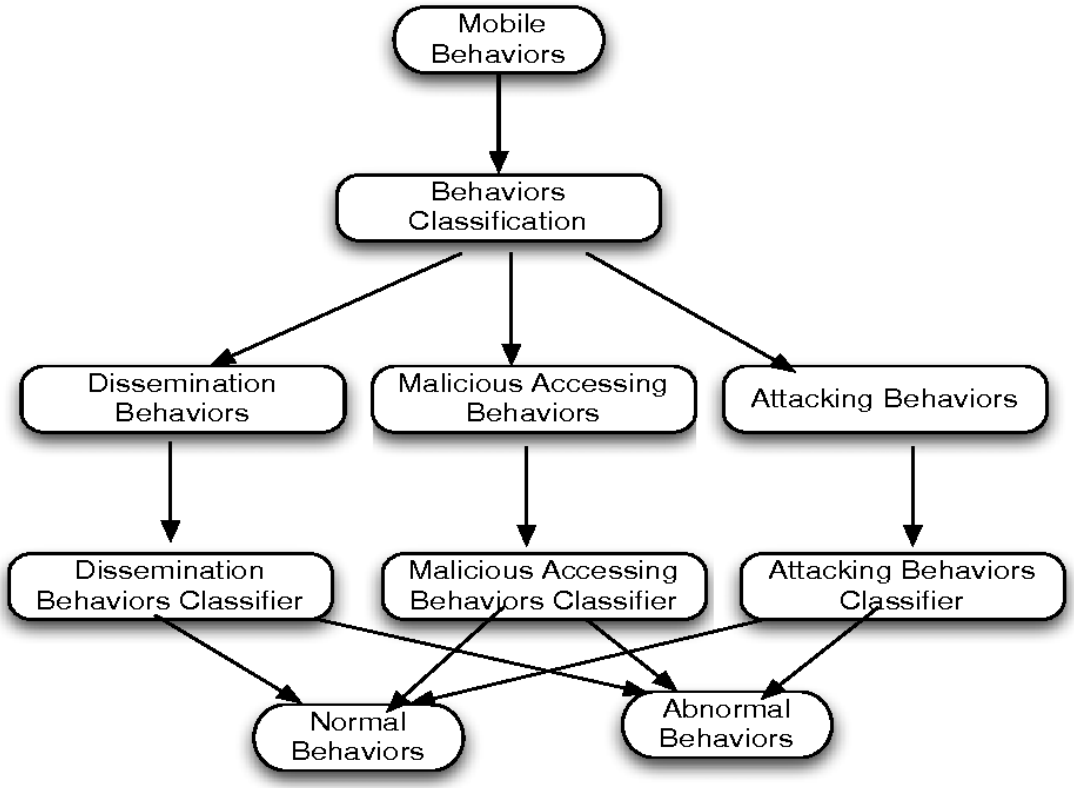

Figure 6. Classifying User Phone Call Behavior

VI. Result and discussion

The findings and analyses from this research project are presented in this section. The time it took to develop the BayesNet model was 0.04 seconds, the Naive Bayes model was built in 0.01 seconds, the Naive Bayes Multinomial Text model was built in 0 seconds, and the Naive Bayes Updateable model was built in 0 seconds, according to the table below.

Table1: Accuracy of the Classifier

SN	**Name of the Classifier**	**Accuracy**
1	NaiveBayes	94.70 %
2	NaiveBayesMultinomialText	84.76%
3	BayesNet	94.67%
4	NaiveBayesUpdateable	94.70%
5	Gaussian Naive Bayes model	95%

VII. Conclusion

A straightforward and adaptable classifier is the Naive Bayes algorithm. The Naive Bayes classifier performs exceptionally well for huge datasets since computation costs are low. In terms of performance, the Naive Bayes classifier outperforms several other classifiers. The Naive Bayes classifier's fundamental assumption of feature independence is one of its main flaws. Features in actual datasets are rarely independent in practice. The Naive Bayes classifier is extremely helpful in the initial interpretation of the data, despite its flaw. Our experimental findings demonstrate that the suggested trust model can identify the malicious entities' strategies for three common attacks: counting-based, time-based, and context-based. Furthermore, unlike the usual approach, which uses simply the global score as a measure of trustworthiness, the suggested trust model learns such strategies as soon as they emerge. Additionally, the recommenders—which are also utilized to accomplish faster and more accurate trust evaluation—resolve the issue of the trust evaluation at the initial interaction. By comparing the Naive Bayes Updateable classifier to existing Bayes classifier models, this research effort suggests the proposed model.

References:

[1]. Jingnian Chen, Houkuan Huang, Shengfeng Tian, and Youli Qu. Feature selectionfor text classi_cation with na□_ve bayes. Expert Systems with Applications,36(3):5432{5435, 2009.

[2]. Bojan Cestnik et al. Estimating probabilities: a crucial task in machine learning. In ECAI, volume 90, pages 147{149, 1990.

[3]. N Eagle, A Pentland, and D Lazer. Infering social network structure using mobile phone data. Proc. of National Academy of Sciences, 2006.

[4]. Luis Daza and Edgar Acuna. An algorithm for detecting noise on supervised classi_cation. In Proceedings of WCECS-07, the 1st World Conference on Engineeringand Computer Science, pages 701{706, 2007.

[5]. Dewan Md Farid, Li Zhang, Chowdhury Mo_zur Rahman, M Alamgir Hossain, andRebecca Strachan. Hybrid decision tree and na□_ve bayes classi_ers for multi-classclassi_cation tasks. Expert Systems with Applications, 41(4):1937{1946, 2014.

[6]. Beno et Fr_enay and Michel Verleysen. Classi_cation in the presence of label noise: asurvey. IEEE transactions on neural networks and learning systems, 25(5):845{869,2014.

[7]. Mark Hall, Eibe Frank, Geo_rey Holmes, Bernhard Pfahringer, Peter Reutemann,and Ian H Witten. The weka data mining software: an update. ACM SIGKDDexplorations newsletter, 11(1):10{18, 2009.

[8]. Martin Halvey, Mark T Keane, and Barry Smyth. Time based segmentation of logdata for user navigation prediction in personalization. In Proceedings of the 2005IEEE/WIC/ACM International Conference on Web Intelligence, pages 636{640.IEEE Computer Society, 2005.

[9]. Jiawei Han, Jian Pei, and Micheline Kamber. Data mining: concepts and tech-niques. Elsevier, 2011.

[10]. George H John and Pat Langley. Estimating continuous distributions in Bayesian classi_ers. In Proceedings of the Eleventh conference on Uncertainty in arti_cialintelligence, pages 338{345. Morgan Kaufmann Publishers Inc., 1995.

[11]. Abhinav Mehrotra, Robert Hendley, and MircoMusolesi. Prefminer: mining

user'spreferences for intelligent mobile noti_cation management. In Proceedings of the2016 ACM International Joint Conference on Pervasive and Ubiquitous Comput-ing, pages 1223{1234. ACM, 2016.

[12]. Mert Ozer, IlkcanKeles, HakkiToroslu, Pinar Karagoz, and Hasan Davulcu. Predictingthe location and time of mobile phone users by using sequential patternmining techniques. The Computer Journal, 59(6):908{922, 2016.

[13]. Veljko Pejovic and MircoMusolesi. Interruptme: designing intelligent promptingmechanisms for pervasive applications. In Proceedings of the 2014 ACM Interna-tional Joint Conference on Pervasive and Ubiquitous Computing, pages 897{908.ACM, 2014.

[14]. Santi Phithakkitnukoon, Ram Dantu, Rob Claxton, and Nathan Eagle. Behaviorbasedadaptive call predictor. ACM Transactions on Autonomous and AdaptiveSystems (TAAS), 6(3):21, 2011.

[15]. J. Ross Quinlan. C4.5: Programs for machine learning. Machine Learning, 1993.

[16]. Iqbal H Sarker, Alan Colman, Muhammad Ashad Kabir, and Jun Han. Behaviororientedtime segmentation for mining individualized rules of mobile phone users.In 2016 IEEE International Conference on Data Science and Advanced Analytics(DSAA), Montreal, Canada, pages 488{497. IEEE, 2016.

[17]. Iqbal H Sarker, Muhammad Ashad Kabir, Alan Colman, and Jun Han. An e_ectivecall prediction model based on noisy mobile phone data. In Proceedings of the2017 ACM International Joint Conference on Pervasive and Ubiquitous Comput-ing: Adjunct. ACM, 2017.

[18]. Vijay Srinivasan, Saeed Moghaddam, and Abhishek Mukherji. Mobileminer: Miningyour frequent patterns on your phone. In ACM International Joint Conferenceon Pervasive and Ubiquitous Computing. ACM, 2014.

[19]. Ian H Witten and Eibe Frank. Data Mining: Practical machine learning tools andtechniques. Morgan Kaufmann, 2005.

[20]. Ian H Witten, Eibe Frank, Leonard E Trigg, Mark A Hall, Geo_rey Holmes, andSally Jo Cunningham. Weka: Practical machine learning tools and techniques withjava implementations. 1999.

[21]. Chia-Chi Wu, Yen-Liang Chen, Yi-Hung Liu, and Xiang-Yu Yang. Decision treeinduction with a constrained number of leaf nodes. Applied Intelligence, 45(3):673{685, 2016.

[22]. Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, HiroshiMotoda, Geo_rey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10algorithms in data mining. Knowledge and information systems, 14(1):1{37, 2008.

[23]. Hengshu Zhu and Enhong Chen. Mining mobile user preferences for personalizedcontext-aware recommendation. ACM Tran. on Intelligent Systems and Technol-ogy, 5(4), 2014.

[24]. Xingquan Zhu and Xindong Wu. Class noise vs. attribute noise: A quantitativestudy. Arti_cial Intelligence Review,22(3):177{210, 2004.

[25]. Mohit Ghandhi " https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c " May5, 2018