International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Mining Iot Data for Next Generation Smart Cities

Abhinav Garg, Manisha Jailia

***Abstract*: Convergence of Cloud, IoT, Networking devices and Data science has ignited a new era of smart cities concept all around us. The backbone of any smart city is the underlying infrastructure involving thousands of IoT devices connected together to work in real time. Data Analytics can play a crucial role in gaining valuable insights into the volumes of data generated by these devices. The objective of this paper is to apply some most commonly used classification algorithms to a real time dataset and compare their performance on IoT data. The performance summary of the algorithms under test is also tabulated**

***Index Terms*: Classification, Data Mining, IoT (Internet of Things), Smart Cities.**

I. INTRODUCTION

Since last few years, there have been enormous discussions around building smart cities (or next generation cities) in India. GOI has proposed to develop few Smart Cities across the country for the renewal, citizen friendly and sustainable development of the country. A smart city is an amalgamation of ICT, networking, Iot, Cloud computing and other Web technologies. It is expected to provide breakthrough solutions to complicated town planning issues, thus improving sustainability and livability." [1]. City planning involves many components like education, infrastructure, transport, municipality, safety, administration etc. Smart cities aim to provide an interconnected, efficient and intelligent living through smart networking and computing technologies like Wireless sensor networks, internet and cloud [2].

Smart cities hold enormous possibilities in turning around the operational competency of a city and IOT (Internet of Things) is the technical basis behind the same. [3]. Internet of Things (IoTs) is an interconnection of various sensing devices and actuators embedded in everyday objects to share the contextual information with the existing applications.

IoT, is an ecosystem of computing devices which are embedded in everyday objects like phones, clock, microwave, doors, switches, shoes etc. Every object has its own unique identifiers (UIDs) and has the ability to transmit data over a network without requiring human intervention [4].

There is an enormous scope of IoT and its applications such as in medicine, wearables, automobiles, agriculture, logistics, smart grids, elderly assistance, retail, healthcare, smart environment, personal, social gaming, robotics, city management and many more. The business value of
Voluminous data produced by the Internet of Things (IoT) is very high. Data Mining algorithms and Knowledge discovery can be applied to this data to discover hidden patterns and reap the actual benefits of large data available. [5]. Information can thereafter be converted into key knowledgeable insights [6]. The results described in [7], [8], [9], [10] shows that data mining algorithms supplements in making IoT smarter for providing smart services in a smart city.

II. IOT DATA MINING ARCHITECTURE

IoT Data Mining Process can be divided in Six stages (Fig 1).

- Stage 1: Sensor and IoT Devices: Interconnected devices in an IoT system application communicate via low power short range wireless communication technologies like Near Field Communication (NFC), WPAN's, SigFox, LoRaWAN, LiFi, ZigBee, iRDA ,Bluetooth and Z-Wave.

-Stage 2: IoT Gateway: Volumes of raw data is collected from the sensors embedded all round us in everyday objects like phones, vehicles, bags, water bottles, clothes, accessories etc. and sent to the gateways which serve as bridge between cloud services and the intelligent devices. An IoT gateway translates between heterogeneous sensor protocols, aggregates, summarizes and preprocesses sensor data and finally forwards it to the cloud storage.

-Stage 3: Cloud Storage: A cloud platform application is built keeping in mind the colossal volumes of data generated by Web, Apps and IoT. It increases scalability and availability of data for real time responses.

-Stage 4: Data Analytics: Data Analytics is a branch of Data Science that makes use of various machine learning tools like neural networks, regression, Artificial Intelligence etc. to identify useful patterns in data. Steps involved in data analysis can be summarized as integration, selection, cleaning, preprocessing, transformation, mining, pattern evaluation and Knowledge presentation. Some data mining techniques are association analysis, classification, clustering, prediction, sequential patterns and decision tree. In this paper we will discuss about the classification algorithms as we focus on finding the efficiency of some well-known classification algorithms on an IoT dataset.

Every learning algorithm behaves differently in different types of problems. Parameters and configurations of the algorithm should be adjusted to attain optimal performance on a dataset.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author

Abhinav Garg*, Assistant Professor, NIFT, Hyderabad, India.
Dr. Manisha Jailia, Associate Professor, Banasthali Vidyapith, Rajasthan , India.

AdaBoost reweighs each training sample to determine the probability of being selected as training set. It aims at converting a number of weak classifiers into a strong one. It is simple to implement and less prone to the overfitting problem than other learning algorithms. [11][12] Refers to AdaBoost (with decision trees as the weak learners) as the best out-of-the-box classifier.

Support Vector machines (SVM) belongs to the category of supervised learning models and works well for binary classification problems and high dimensional data. Face/Image/Handwriting detection, Bioinformatics to name a few. High algorithmic complexity and extensive memory required for SVM make it a bad choice for wireless sensor networks.

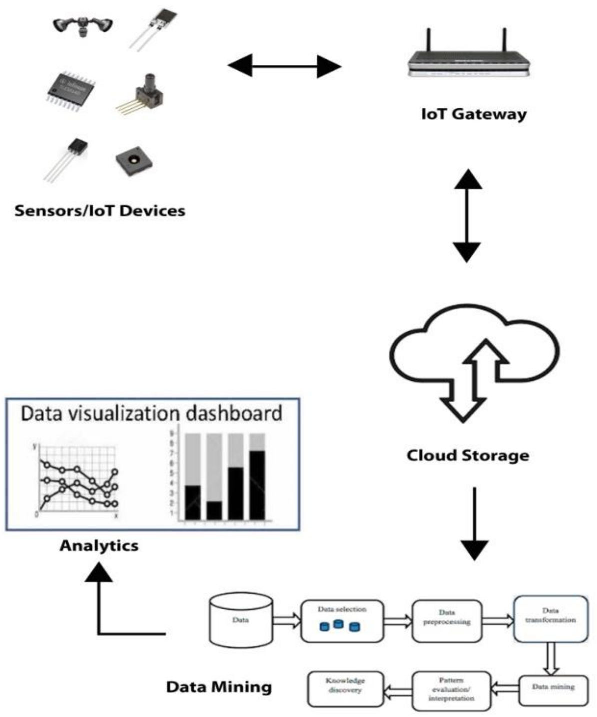
Decision Tree algorithms like CART, ID3,C4.5 are another supervised learner that are easy to visualize and works well with both numerical as well as categorical data, but as the tree grows larger ,it over fits the training data and a small change in data might rearrange the entire tree.

Neural Networks are built around the concept of human brain(neurons and perceptron). Results of Neural networks and SVMs are highly accurate and nearly same. Comparatively lesser training and running time enables SVMs to be scaled to large data sets, so they can be used to replace neural networks. However according to [19] training step of SVMs is slow and requires huge computing resources in comparison to other classification algorithms. Contrarily, neural networks have already proven their worth in multi-category classifications and many other applications such as dimension reduction, anomaly detection, regression etc. [20].

Naiye Bayes is a probabilistic classifier which is easily constructible and scalable to huge data sets [13]. Naïve Bayesian classifier gives promising results in terms of categorical classification accuracy, in spite of conditionally independent features for a given class label and is popularly used in healthcare, recommender systems, spam filtering etc. [21] Domb stated that performance of Random forest controls overfitting and is better than the decision tree algorithms in most of the cases. But complex algorithm and difficult implementation make it a bad choice for real time IOT applications with limited memory space and processing capabilities [22]. KNN classifiers are no so quick, so they lie in the category of lazy learners. K-nearest neighbors algorithm works on the principle of calculating the distances between labelled and unlabeled data in a dataset. In KNN building the model is cheap, but computation cost is high as classifying new cases on the basis of similarity measure is comparatively expensive [13]. KNN has been used extensively in wireless sensor networks (WSNs) and IoT domain for recommender systems, Pattern recognition, intrusion detection [14], statistical estimation, indoor positioning systems [15], transaction-scrutinizing software applications and activity recognition [16].

In some cases kNN have outperformed SVM. [17]. At some places, kNN classifier that uses DTW as a similarity measure outperformed other conventional classification techniques like NN, kNN, SVM while conventional KNN (with parameter setting of K = 1, 5, 10) is ranked lowest among the classification approaches [18].

- Stage 5: Visualization and Insights: Data visualization is the graphical representation of information or patterns in form of easy to interpret colorful charts, tables, maps, infographics, dashboards and display. It provides deep insights into the dataset and helps in better decision making



. Fig 1: Data Mining Architecture

III. EXPERIMENTATION AND METHODOLOGIES

Classification is a supervised data mining technique in which a set of labelled data (Training set) is used to categorize new observations. It is widely used in spam detection, sentiment analysis, machine learning, speech recognition etc.

For evaluating the performance of various data mining classification algorithms we have taken a dataset that predicts the state of person (running / walking) from the data collected by sensors in different iOS devices placed in the vicinity. Accelerometer and gyroscope of iPhone collected 88588 samples of data at an interval of 10 seconds ~5.4/second frequency [23].

Sensor data:

Input variables - acceleration_x, acceleration_y, acceleration_z, gyro_x, gyro_y, gyro_z

Output Variable – (Class Label): "activity" where “0" represents walking and “1" represents running.

All the experimentation and workflows are designed on a machine learning and data visualization tool Orange version 3.8. (Fig. 2) Orange features visual programming front end for interactive data visualization. Before simulating the algorithms the datasets were preprocessed to remove any outliers or missing values.

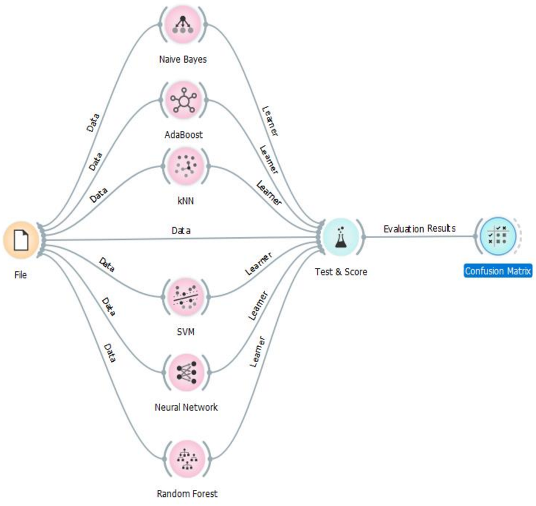
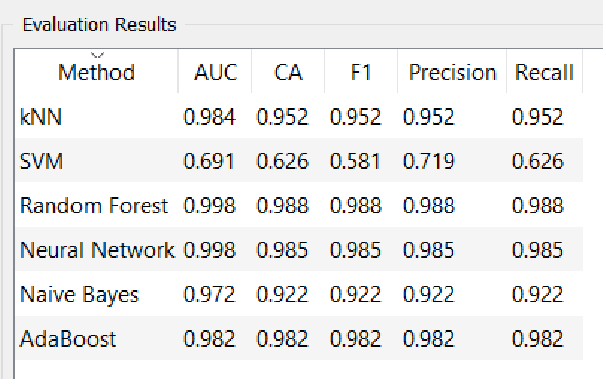


Fig 2: Workflow diagram of experiment

IV. EVALUATION

Six well known classification algorithms Random Forest, Naive Bayes, SVM, Ada boost, kNN, Neural Networks were used in predicting the target class ‘Activity’(Walk/Run). It was observed that classification accuracy of Random Forest, Ada Boost and Neural Networks is almost comparable whereas kNN and Naive Bayes gave a slightly lower accuracy rates. In our experiment SVM depicted relatively lower classification accuracy than its counter parts. To describe the performance of classification models, evaluation matrix and confusion matrix of all the learners under given test data is given in Fig 3 and 4 respectively.



Evaluation Results

Method	AUC	CA	F1	Precision	Recall
kNN	0.984	0.952	0.952	0.952	0.952
SVM	0.691	0.626	0.581	0.719	0.626
Random Forest	0.998	0.988	0.988	0.988	0.988
Neural Network	0.998	0.985	0.985	0.985	0.985
Naive Bayes	0.972	0.922	0.922	0.922	0.922
AdaBoost	0.982	0.982	0.982	0.982	0.982

Fig. - 3

A binary classifier divides the test data in number of True Negatives (TN), True Positives (TP), False Negative (FN), and False Positives (FP). The performance of algorithms under consideration here is given in terms of:

- AUC (Area under Curve)
- Classification Accuracy

(CA) = (TP + TN) / (P + N)

- Precision (Positive predicted value)

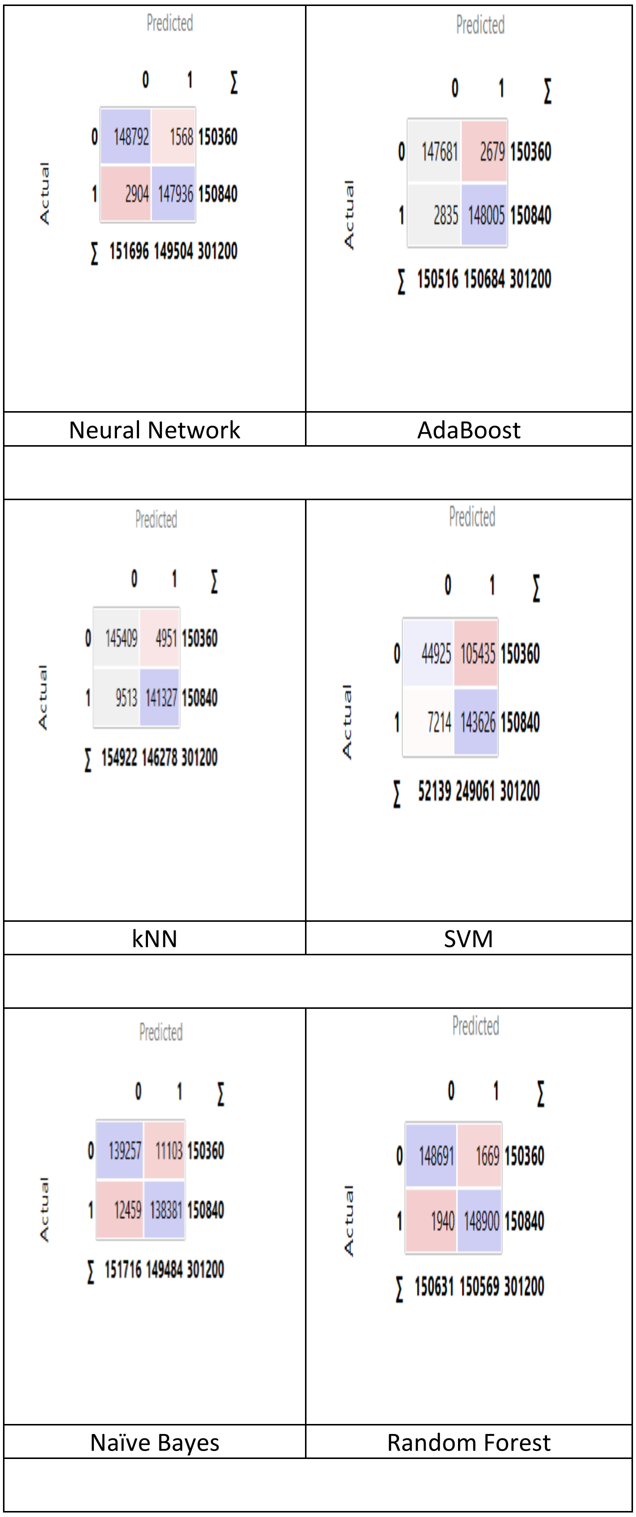
= TP / (TP + FP)

- F1 (harmonic mean of precision and sensitivity)

= TP / (2TP + FP + FN)

- Recall (Actual Positives that are correctly labeled)

= TP / (TP + FN).



Neural Network

Actual \ Predicted	0	1	∑
0	148792	1568	150360
1	2904	147936	150840
∑	151696	149504	301200

AdaBoost

Actual \ Predicted	0	1	∑
0	147681	2679	150360
1	2835	148005	150840
∑	150516	150684	301200

kNN

Actual \ Predicted	0	1	∑
0	145409	4951	150360
1	9513	141327	150840
∑	154922	146278	301200

SVM

Actual \ Predicted	0	1	∑
0	44925	105435	150360
1	7214	143626	150840
∑	52139	249061	301200

Naïve Bayes

Actual \ Predicted	0	1	∑
0	139257	11103	150360
1	12459	138381	150840
∑	151716	149484	301200

Random Forest

Actual \ Predicted	0	1	∑
0	148691	1669	150360
1	1940	148900	150840
∑	150631	150569	301200

Fig. 4:- Confusion matrix for a) Neural Network b) KNN c) Naive Bayes d) AdaBoost e) SVM f) Random Forest

V. CONCLUSIONS

The dream of smart cities and a smarter living has generated thousands of connected-intelligent devices and open source tools to manage data that can help us in providing insights to change customer experiences, drive unprecedented efficiencies, and develop new products / business models.

Retrieval Number E7537068519 /2019©BEIESP
DOI: 10.35940/ijeat.E7537.088619
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The success of Smart City Mission in India will be mainly based on the seamless, coherence and interoperable execution of the IoT's involved.

The data set used in our experiment is comparatively smaller than the actual deluge of data produced by numerous ICT devices in connection. However ever increasing large volumes of information generated by Wireless Sensor networks and ICT devices calls for Smarter Data Science and technologies.

REFERENCES

1. Toppeta, D. (2010). The Smart City Vision: How Innovation and ICT Can Build Smart, "Livable", Sustainable Cities. The Innovation Knowledge Foundation. Available from http://www.thinkinnovation.org/file/research/23/en/Top
2. peta_Report_005_2010.pdf (accessed on 26th December 2018).
3. Washburn, D., Sindhu, U., Balaouras, S., Dines, R. A., Hayes, N. M., & Nelson, L. E. (2010). Helping CIOs Understand "Smart City" Initiatives: Defining the Smart City, Its Drivers, and the Role of the CIO. Cambridge, MA: Forrester Research, Inc. Available
4. From http://public.dhe.ibm.com/partnerworld/pub/smb/smarterplanet/forr_help_cios_und_smart_city_initiatives.pdf (accessed on 26th December 2018).
5. IoT to Play a Major Role in building a Smart City. Available at:- https://internet-ofthings.cioreviewindia.com/cioviewpoint/iot-to-play-a-major-role-in-building-a-smart-city-nid-689-cid1.html (accessed on 26th December 2018).
6. Internet of Things available at https://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT (accessed on 26th December 2018).
7. F. Chen, P. Deng, J. Wan, D. Zhang, A. V. Vasilakos, and X. Rong, "Data mining for the internet of things: literature review and challenges," International Journal of Distributed Sensor Networks. vol. 2015, no. 9, 2015, Art. ID 431047.
8. C. Tsai, C. Lai, M. Chiang, and L. T. Yang, "Data mining for Internet of Things: A survey," IEEE Commun. Surveys Tuts., vol. 16, no. 1, pp. 77– 97, 1st Quart. 2014.
9. V. Cantoni, L. Lombardi, and P. Lombardi, "Challenges for data mining in distributed sensor networks," in Proc. International Conference on Pattern Recognition, vol. 1, 2006, pp. 1000–1007.
10. T. Keller, "Mining the internet of things: Detection of false-positive RFID tag reads using low-level reader data," Ph.D. dissertation, The University of St. Gallen, Germany, 2011.
11. E. Masciari, "A framework for outlier mining in RFID data," in Proc. International Database Engineering and Applications Symposium, 2007, pp. 263–267
12. S. Bin, L. Yuan, and W. Xiaoyi, "Research on data mining models for the internet of things," in Proc. International Conference on Image Analysis and Signal Processing, 2010, pp. 127–132.
13. B. Kégl, The return of AdaBoost. MH: multi-class Hamming trees, arXivpreprint arXiv:1312.6086.
14. Joglekar, S. Adaboost—Sachin Joglekar's Blog. Available online: https://codesachin.wordpress.com/tag/adaboost/ (accessed on 26th December 2018).
15. X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. F. M. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," Knowl. Inf. Syst., vol. 14, no. 1, pp. 1–37, 2008.
16. W. Li, P. Yi, Y. Wu, L. Pan, and J. Li, ''A new intrusion detection system based on KNN classification algorithm in wireless sensor network,'' J. Elect. Comput. Eng., vol. 2014, Jun. 2014, Art. no. 240217.
17. D. Li, B. Zhang, and C. Li, ''A feature-scaling-based k-nearest neighbor algorithm for indoor positioning systems,'' IEEE Internet Things J., vol. 3, no. 4, pp. 590–597, Apr. 2016.
18. Serge Thomas, Mickala Bourobou, Younghwan Yoo, "User Activity Recognition in Smart Homes Using Pattern Clustering Applied to Temporal ANN Algorithm", Sensors, vol. 15, pp. 11953-11971, 2015.
19. Kuramochi, M., and Karypis, G., "Gene classification using expression profiles: a feasibility study", International Journal on Artificial Intelligence Tools, 14 (4) (2005) 641-660.
20. A. Azmoodeh, A. Dehghantanha, M. Conti, and K.-K. R. Choo, "Detecting crypto-ransomware in iot networks based on energy consumption footprint", Journal of Ambient Intelligence and Humanized Computing, 2017.
21. I. Yoo, P. Alafaireet, M. Marinov et al., "Data mining in healthcare and biomedicine: a survey of the literature," Journal of Medical Systems, vol. 36, no. 4, pp. 2431–2448, 2012.
22. Mukkamala S., Janoski G., Sung A. H. (2002) "Intrusion Detection Using Neural Networks and Support Vector Machines," Proceedings of IEEE International Joint Conference on Neural Networks, pp.1702-1707.
23. Kononenko, I., Machine learning for medical diagnosis: history, state of the art and perspective. Artif. Intell. Med.23:89–109, 2001.
24. Domb M., "An Adaptive Lightweight Security Framework Suited for IoT" available at https://www.intechopen.com/books/internet-of-things-technology-applications-and-standardization/an-adaptive-lightweight security-framework-suited-for-iot (accessed on 26th December 2018).
25. Sensor data available at https://www.openml.org/d/40922 (accessed on 26th December 2018).

AUTHORS PROFILE

Abhinav Garg is Assistant Professor in Fashion Communication Deptt at NIFT Hyderabad..
He has done B.Tech. (C.S.E.) and M.Tech. (I.T.)
His areas of Interest Include Data Mining, Intellectual Property, Design Management and CAD.

Dr. Manisha Jailia is Associate Professor in Computer Science Deptt. at Banasthali Vidyapith, Rajasthan India. She has done MCA, UGC-NET, and Ph.D.
Her research areas are Data Mining, Distributed Databases, Web technologies, Big Data, Cloud Computing, and High-performance computing.