ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20220305125327

Preliminary communication
Received: 2022-03-05, Accepted: 2022-03-14

An Enhanced AdaBoost Classifier for Smart City Big Data Analytics

Fahd S. Alotaibi

Abstract: The targeted goal regarding the smart cities is improving the goodness of their people and to raise the economic improvement in maintaining certain rate or level. Smart cities would increase all set of utilities, which involves healthcare, education, transportation and agriculture among other utilities. Smart cities are depended on the ICT framework, which includes the Internet of Things methodology. These methodologies make bulk of diverse in data, which referred to as big data. Moreover, these data have no purpose by themselves. Modules needed to improve as new to explain the large amount of data collected and one of the good methods to solve is to use the methods of big data analytics. It shall be maintained and designed through the methods of analytics to get good understanding and in order to increase the utilities of smart city.

Keywords: AdaBoost; Big Data; Internet of Things; Linear Regression; Smart Cities

1 INTRODUCTION

Cities all over the world are trying to change themselves into smart cities. Most recent research shows that the main factor in this change is urban big data use from the things, which are physical in city areas. The usage of data in smart cities constantly remains Strange by framework and knowledge. This research paper results in finding an analysis on cases, which are different types among big data in cities of all over the world and government organizations projects toward smart cities development. [1] The data use for smart cities can form a framework by collecting the models for reference, problems to be faced, and thoughts. Generating huge amount of data in different format and takes from more parts like traffic sector, energy sector, education sector, healthcare and producing various parts is the main application in the smart city. [2] The produced data is gathered in huge amounts and on general it offers a view on what and how was happening in real-time of the city at any time. To confirm the correction and needful using these data in applications of smart city, which are perfectly suit and powerful tools among big data management must be present. [3]

2 IMPACTS OF BIG DATA

Impacts of Big Data includes different departments such as transport, cost, safety etc. [4] (Fig. 1).

1) **Public Safety.** Identity the prone area for the purpose of public safety to predict the exact crime location.
2) **Transportation.** Traffic jam and roadblock can be decreased and road optimization can be done by data driven.
3) **Cost Minimization.** Used to identify the required area transformation and identity what kind of transformation.
4) **Supportable Growth.** Growth drivers of suitability is by Continuous growth. The outcome in development of a smart city is determined by the major playing role of data.
5) **Smart Network Infrastructure.** It contains the capacity of connecting components easily. The real-time smart cities applications in big data must have the support of quality of service (QoS).
6) **Smart Filtering and Aggregating.** It helps to decrease the traffic of network and fastest data preprocessing.

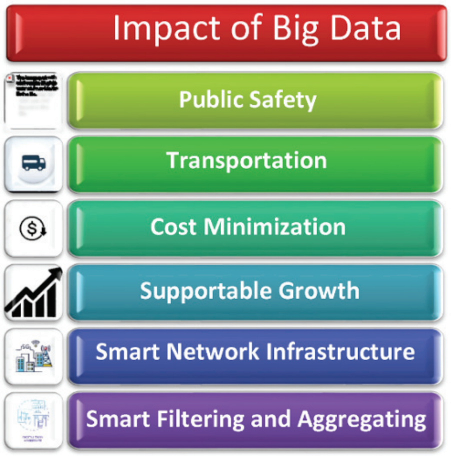


Figure 1 Big Data Impacts

3 ADABOOST CLASSIFICATION ALGORITHM

Methods for classification must have developed for decades. There are two types of classification methods commonly termed as: supervised classification and unsupervised classification. The analyst chooses the training samples land cover class first for each and it makes guidance for the system to find same areas in each class is known as Supervised Classification. From The selection of given trained samples which is depended on collection of field data. The best and recent classification methods of supervised machine learning includes large likelihood method, parallelepiped method, small distance, decision tree method, random forest method, and support vector machine method, among other methods. [5] The classification, which does not start with training samples, is known as unsupervised classification. However, the analyst particularly choose the desired count of classes, and hence the system automatically groups the pixels, which are closely same as the clustering algorithms. [6]

The cluster algorithms, which are commonly used, are K-Means, Iterative Self-Organizing Technique for Data Analyzing. The Repetitive process of cluster produces in a preset count of "spectral classes", which then declared as labels of class and changes to "information classes". Unsupervised is the method of classification, which is especially efficient when, came to know early knowledge about the research area, which was unavailable? Sometimes hybrid approach is used which merges the unsupervised and supervised classification methods. [7]

To classify remote images supervised and unsupervised methods are used. They are the classification methods, which are pixel-based, which is only depended on information of spectral. [8]

3.1 AdaBoost

Adaptive Boosting is shortly known as AdaBoost Algorithm, which is a technique for boosting used in Machine Learning. It is called Adaptive Boosting Algorithm as the weights were assigned again for each instance, with heavier weights assigned with incorrectly categorized instances. Decreasing as well as growing difference in supervised machine learning is done boosting. It does something on learners growing principle Back-to-Back. Each ensuring learner is grown from previously grown learners Except the first. Shortly, weak learners were changed as strong learners. The algorithm of AdaBoost tries as similar as the rule for boosting it with a little variance. [9]

During the period or hour of data training [10], it makes *n* count of decision trees. As the first decision tree or sample demo is made, the first model priority is given to incorrect records, which are classified. As input for the second model priority, only these records are sent. Until we denote a number of base learners, the process goes on and we need to make creation. Point to remember, with all boosting techniques repetition of records are allowed.

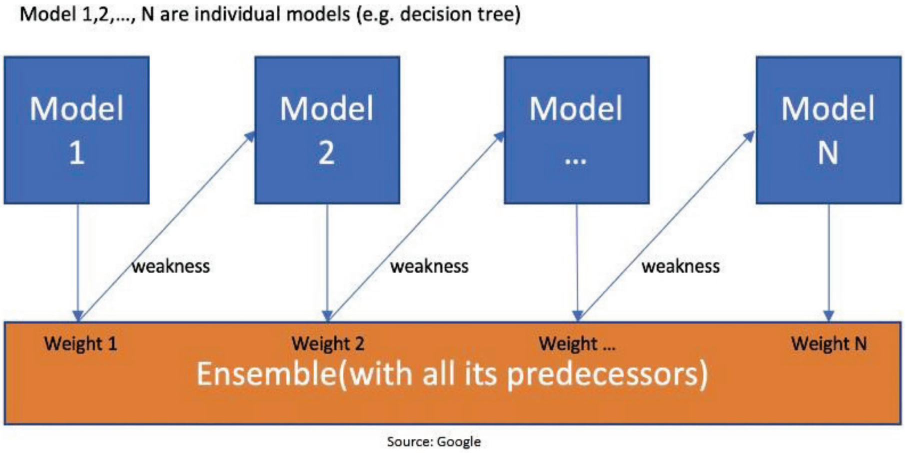


Figure 2 Work Structure of AdaBoost Classifier

4 IMPLEMENTATION OF ADABOOST CLASSIFIER ALGORITHM WITH PYTHON

In AdaBoost Classification [14] To the data points, higher points are allotted which are not classified properly or predicted wrongly by the previous model. This determines that a weighted input will be got through each successive model.

The AdaBoost model consists of weak classifiers, weight update and classify.

Weak Classifiers. AdaBoost combines weak classifiers with certain strategies to get a strong classifier, as shown below. At each iteration, the weights of samples, which are wrongly classified, will increase to catch the classifier "attention". For example, in Fig. 3a, the dotted line is the classifier-plane and there are two blue samples and one red sample, which are wrong, classified. Then, in Fig. 3b, the weights of two blue samples and one red sample are increased. After adjusting the weights at each iteration, we can combine all the weak classifiers to get the final strong classifier.

Weight Update. There are two types of weight to update at each iteration, namely, the weight of each sample and the weight of each weak classifiers. At the beginning, there are initialized as follows:

$$w_i = \frac{1}{n} \tag{1}$$

$$\alpha_m = \frac{1}{m} \tag{2}$$

where *n* and *m* are the number of samples and the number of weak classifiers respectively.

AdaBoost trains a weak classifier at each iteration denoted as whose training error is calculated as:

$$e_m = \sum\nolimits_{i=1}^{n} w_{mi} I(G_m(x_i)) \neq y_i \tag{3}$$

Then, update the weight of weak classifier by:

$$\alpha_m = \frac{1}{2}\ln\left(\frac{1-e_m}{e_m}\right) \tag{4}$$

From the above equations, we can conclude that:

1) The training error is the sum of weights of the wrong classified samples.
2) When e_m is less than 0.5, α_m is greater than 0, which means the lower training error the weak classifiers has, the more important role that weak classifier plays in the final classifier.
3) The code of training process [11-14] of AdaBoost is shown below:

```
def train (self, train_data, train_label):
    if self.norm_type == "Standardization":
train_data = preprocess.Standardization(train_data)
```

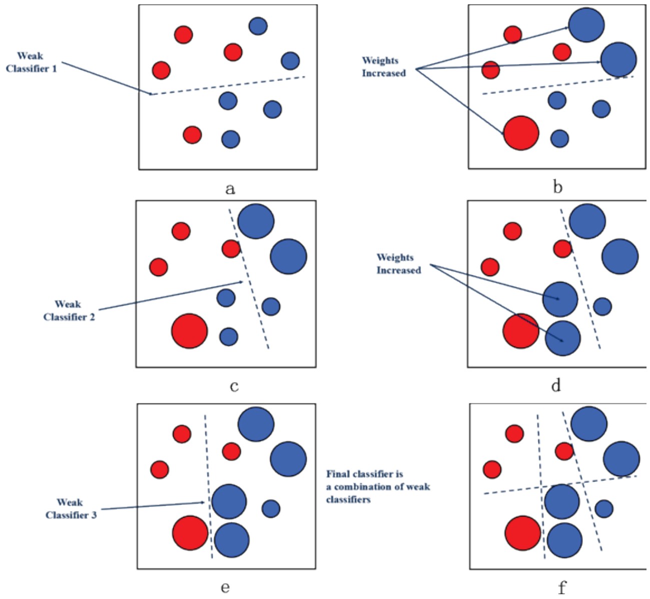


Figure 3 Week and Strong Learners of AdaBoost Classifier AdaBoost Model

Conclusion and Analysis of AdaBoost Implementation. AdaBoost can be regarded as additive model with exponent loss function using forward step algorithm. In AdaBoost, the type of weak classifiers can be different or the same [15]. In this article, we use 5 SVM classifiers as the weak classifiers, and the detection performance is shown below:

```
Output:
Accuracy of AdaBoost:   0.850000
Runtime of AdaBoost:    3. 4339-06546173096
```

AdaBoost Algorithm:

Step 1: Assign Entire observations with its own weight

Step 2: With the support of stump the observation samples are classified randomly

Step 3: Sum of weights of misclassified record is calculated in the form of total error.

Total Error = weight of misclassified records Total error will among 1 and 0.

1 represents misclassification which is also known as weak stump.

0 represents correct classification which is also termed as perfect stump.

Step 4: calculate stump performance

$$SP = \frac{1}{2}\ln\left(\frac{1-TE}{TE}\right) \tag{5}$$

(Where SP is stump performance as well as TE is total error)

Step 5: Update weight based on the performance of stump,the weights are updated

$$\begin{aligned} NW &= W * E^{P} > MR \\ NW &= W * E^{-P} > CCR \end{aligned} \tag{6}$$

Where NW is new weight, W is weight, P is performance, MR is misclassified records, and CCR is correctly classified records.

Step 6: Update weights in Iteration.

Step 7: Discover Final Predictions

$$\frac{FP}{SIGN\,(WS)} = \sum(\alpha_i * EIWPV) \tag{7}$$

Where FP is final prediction, WS is weighted sum and EIWPV is each iteration with predicted value.

Linear Regression. While we have a single input in Linear Regression, to show the coefficients efficiently statistics can be used.

Calculating statistical properties is essential from the data that are:

- means,
- standard deviations,
- correlations and
- covariance.

To cross and calculate the method of statistics, all data must available.

5 RESULT ANALYSIS

Smart city index is the dataset from kaggle.com

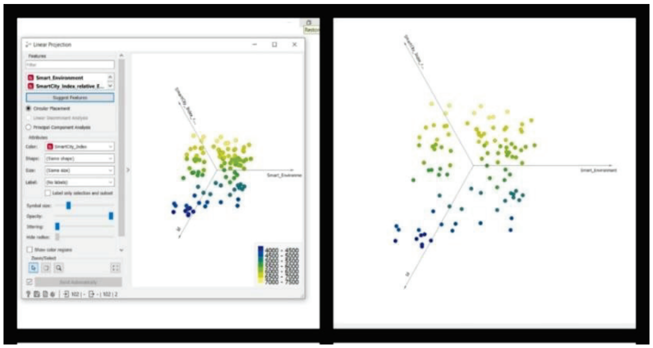
Utilizing different types of IoT (Internet of Things) sensors to collect and manage data - combined with many other technical integrations into our city hubs - defines the future of data & automation being embedded in our urban-living. Think of Smart Cities as a customer experience - for residents of a city.

The Leap Data team utilized globally recognized indices (formalized for the evaluation of Smart City initiatives), and developed a data model to interpret how Calgary & Edmonton stand in relation to Global Leaders of Smart City activities.

6 ATTRIBUTES

Table 1 Smart City Index

ID	City	Country	Smart Mobility	Smart Mobility	Smart City Index
1	Oslo	Norway	6480	6512	7138
2	Bergen	Norway	7097	6876	7296
3	Amsterdam	Netherlands	7540	5558	7311
4	Copenhagen	Denmark	7490	7920	7171
5	Stockholm	Sweden	6122	7692	6812
6	Montreal	Canada	7490	4848	7353
7	Vienna	Austria	5683	7608	6771
8	Odense	Denmark	6160	8404	6886
9	Singapore	Singapore	5790	4344	6813
10	Boston	United States	7870	5224	6852



Linear Projection

Features

Smart_Environment

SmartCity_Index_relative_E...

○ Circular Placement

○ Linear Discriminant Analysis

○ Principal Component Analysis

Attributes

Color: SmartCity_Index

Shape: (Same shape)

Size: (Same size)

Label: (No labels)

□ Label only selection and subset

□ Show color regions

Figure 4 Outcome of AdaBoost

ID: Column identifier

City: List of smart cities across the world

Country: List of countries where smart cities across the world are located

Smart Mobility: Index calculated from assessment of citywide Public Transportation System, ICT, and accessibility infrastructure.

Smart Environment: Index calculated from environmental sustainability impact, monitoring pollution and energy management.

Smart Government: Index calculated from comparative study of transparent governance & open data initiatives of smart cities across the Government of each States.

Smart Economy: index calculated through global comparison of citywide productivity, economic vitality, and support for Finance.

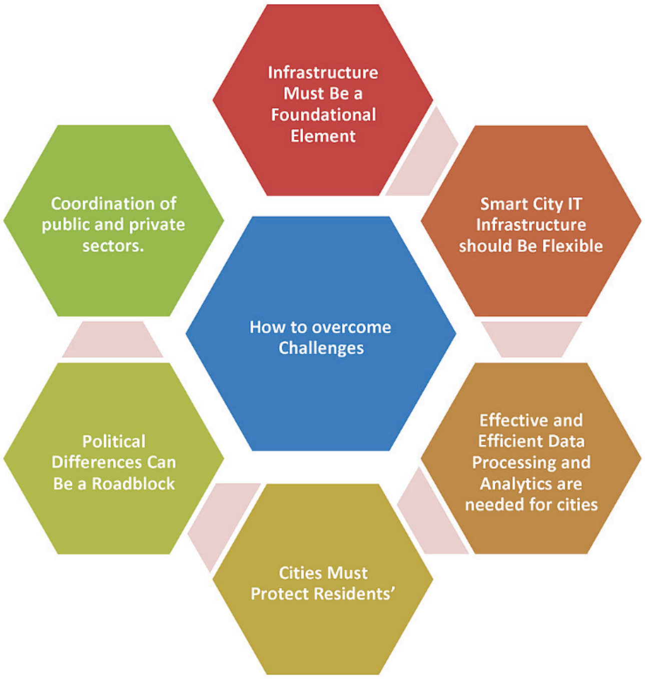
Smart People: Index calculated by comparing social and cultural plurality, education systems and its supporting ancillary facilities across the world.

Smart Living: Index calculated by measuring metric around healthcare services, social security and housing quality.

Smart City Index: Aggregate score for smart city model based on smart city super groups.

7 CHALLENGES AND OVERCOMING TECHNIQUES

Smart city planning is the combination of various sectors that adds peoples, public welfare organizations, local and state government and private enterprises, healthcare, etc. hence after these combinations are gathered, it develops various large chances for business, sustainability, disaster prevention, public safety. However, many difficulties and problems can be shown by the combination of methodological collaboration and innovation among the private enterprises and public organizations.



Infrastructure Must Be a Foundational Element

Coordination of public and private sectors.

Smart City IT Infrastructure should Be Flexible

How to overcome Challenges

Political Differences Can Be a Roadblock

Effective and Efficient Data Processing and Analytics are needed for cities

Cities Must Protect Residents'

Figure 5 Methods to overcome Challenges

8 CONCLUSION

Two important concepts are Big Data and Smart City; Hence, Applications in developing Smart Cities will help reaching the comfortability, good recovery than previous, Powerful manner of governance, Increased Life's Quality, and brilliant Organization of resources in smart city Development. Our work of research explained either the terms or their various short Explanation and then we were gone to know that some attributes which are common for everything. In spite of differing the short explanation, each category has a count of characteristics and functions that particularly means it. Based upon the above functionalities, which are common, hence now we are capable of choosing the most similar gain of big data usage to make design and support the improvement of applications in smart city.

9 REFERENCES

[1] Niu, A., Cai, B., & Cai, S. (2020). Big data analytics for complex credit risk assessment of network lending based on SMOTE algorithm. *Complexity*, 2020. https://doi.org/10.1155/2020/8563030

[2] Joseph, L. L., Goel, P., Jain, A., Rajyalakshmi, K., Gulati, K., & Singh, P. (2021, October). A Novel Hybrid Deep Learning Algorithm for Smart City Traffic Congestion Predictions. In *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), IEEE*, 561-565. https://doi.org/10.1109/ISPCC53510.2021.9609467

[3] Rathore, M. M., Son, H., Ahmad, A., & Paul, A. (2018). Real-time video processing for traffic control in smart city using Hadoop ecosystem with GPUs. *Soft Computing, 22*(5), 1533-1544. https://doi.org/10.1007/s00500-017-2942-7

[4] Wu, D. (2019, January). An audio classification approach based on machine learning. In *2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), IEEE*, 626-629. https://doi.org/10.1109/ICITBS.2019.00156

[5] Sulaiman, M. A. (2020). Evaluating Data Mining Classification Methods Performance in Internet of Things Applications. *Journal of Soft Computing and Data Mining, 1*(2), 11-25.

[6] Mani, J. J. (2021). Predictive Modeling Framework for Diabetes Classification Using Big Data Tools and Machine Learning. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12*(10), 818-823. https://doi.org/10.17762/turcomat.v12i10.4255

[7] Banga, A., Ahuja, R., & Sharma, S. C. (2022). Stacking regression algorithms to predict PM2.5 in the smart city using internet of things. *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 15*(1), 60-76. https://doi.org/10.2174/2666255813999200628094351

[8] Niu, A., Cai, B., & Cai, S. (2020). Big data analytics for complex credit risk assessment of network lending based on SMOTE algorithm. *Complexity*, 2020. https://doi.org/10.1155/2020/8563030

[9] Munawir, H., Mabrukah, P. R., & Djunaidi, M. (2021). Analysis of green supply chain management performance with green supply chain operation reference at the batik enterprise. *Economic Annals-XXI*, 187. https://doi.org/10.21003/ea.V187-14

[10] Joseph, L. L., Goel, P., Jain, A., Rajyalakshmi, K., Gulati, K., & Singh, P. (2021, October). A Novel Hybrid Deep Learning Algorithm for Smart City Traffic Congestion Predictions. In *The 6th International Conference on Signal Processing, Computing and Control (ISPCC2021)*, IEEE, 561-565. https://doi.org/10.1109/ISPCC53510.2021.9609467

[11] Bidmeshki, G. A. & Taheri, F. (2018). Investigating the Effect of Emotional Intelligence on Job Performance (Case Study: Employees of Islamic Azad University, Qaemshahr Branch). *Journal of Management and Accounting Studies, 6*(02), 33-38. https://doi.org/10.24200/jmas.vol6iss02pp33-38

[12] Shirvani, M., Mohammadi, A., & Shirvani, F. (2015). Comparative study of cultural and social factors affecting urban and rural women's Burnout in Shahrekord Township. *Journal of Management and Accounting Studies, 3*(01), 1-4.

[13] Wang, F., Jiang, D., Wen, H., & Song, H. (2019). Adaboost-based security level classification of mobile intelligent terminals. *The Journal of Supercomputing, 75*(11), 7460-7478. https://doi.org/10.1007/s11227-019-02954-y

[14] Habibzadeh, H., Kaptan, C., Soyata, T., Kantarci, B., & Boukerche, A. (2019). Smart city system design: A comprehensive study of the application and data planes. *ACM Computing Surveys (CSUR)*, 52(2), 1-38. https://doi.org/10.1145/3309545

Author's contacts:

Fahd S. Alotaibi
Information Systems Department,
Faculty of Computing and Information Technology,
King Abdulaziz University,
Jeddah, Saudi Arabia
E-mail: fsalotaibi@kau.edu.sa