A Clustering Hybrid Algorithm for Smart Datasets using Machine Learning

Dar Masroof Amin[1]
Research Scholar
MMICT and BM Maharishi Markandeshwar
(Deemed to be University)
Mullana, Haryana 133203, India

Dr. Munishwar Rai[2]
Professor
MMICT and BM Maharishi Markandeshwar
(Deemed to be University)
Mullana, Haryana 133203, India

***Abstract*—In the field of data science, Machine Learning is treated as sub-field which primarily deals with designing of algorithms which have ability to learn from previous information and make future predictions accordingly. In traditional computational world the Machine Learning was generally performed on highly performance servers and machines. The implementation of these concepts on Big Data analytics algorithms has high potential and is still in its early stages. So far as machine learning is concerned, performance measure is an important parameter to evaluate the overall functionality of the algorithms. The data set is a different entity and the measuring of performance on a data which is unseen is also called as test set, and training set is a Data set which is training itself. The Data Mining is extensively using learning algorithms for data analysis and to formulate future predications based on archived data. The research presented provides a step forward to make smart data sets out of training data set by evaluating machine learning algorithm. The research presented a novel hybrid algorithm that attempts to incorporate the feature of similarities in Random Forest machine learning algorithm for improving the classification accuracy and efficiency of working**.

Keywords---Random Forests (RF); Jaccard Similarity (JS); triangle; smart data; root mean square error; mean absolute error; machine learning

I. INTRODUCTION

Big Data terminology is generally applied to the data that grows exponentially and which cannot be accessed by using conventional database systems. The size of data sets involved in big data cannot be handled by traditional software technology and database. The common tools, storage systems cannot store, process and manage the size of datasets [1]. The big data analytics is changing the overall life on the globe in various aspects, viz. health care, marketing, etc. [2]. The new technology and techniques are getting imbibed into out day to utility, Internet of Things (IoT) devices. The technologies being used generates valuable data which can in turn be used for making important decisions [3]. The data that is generated out of these devices can be either result of conscious intervention or unintentional. This involvement of the human in creating of data is at same time creating opportunities for analyzing the data for various purposes. The number of devices that were connected to internet and were generating the data was double in 2008 than general human beings. The expectation is that it will reach up to 50 billion by the end of 2020 and hence the creation of data can be seen as exponentially growing [4]. In a similar manner by 2025 the economy is like to grow $11.1 trillion a year [5]. This is the reason that multinational corporations' are moving towards big data technologies to improve their skills for making additional profit out of their investments [6]. The increase in general technology is also playing a vital role in meeting the demands of growing data [7]. The solution provided by the IoT is by combining the information technology with hardware and software. The Big Data analytics generally provides the soft solutions to handle the exponential growth of data. The physical functionality of digital devices is then accessed locally and globally [8]. The creation of human sense while providing soft solutions to the problems arising out of growth of data, the cost of security system and other functions can be minimized at optimal functionality. The application of distributed system in data analytics will minimize the load over the system. The data generated from IoT should be explored for using it in a formal process. As the traditional analysis of data was providing reports or models on the basis of data available in the system similarly the Big Data coming out of IoT is providing analysis of data generated on real time basis. The system should handle this real time input properly and should provide an optimal solution to the user of the data so far as decisions are concerned. This analytics enables smart decision making and means of quantification and goal tracking. And the traditional modulation provides analysis of static data analysis of unstructured data [9]. The case of Big Data is complex where large data is involved and which needs finding correlations between various types of input in real time. In outmoded analysis of data, the archival data is used for extracting and establishing relationship among various variables. Machine learning instead begins with the result of variables involved and thereon uses the interaction of the predictor variables. The Google's machine learning application is reducing the use of energy and making cool environment for their data center. By adopting machine learning technology, Google will save millions of dollars in creating a favorable environment for their servers. The machine learning algorithms have capability to learn various physical environmental changes and looking hidden patterns in data and later on making smart decisions to tackle any odd situation. The use of the technology have improved the services of smart homes, healthcare, agriculture etc. In the upcoming years billions of machines and devices will be connected to the internet and therefore data generated will be huge. The gigantic growth in the data have to be tackled by

smart machine learning algorithms in order to reap the overall benefit of this growing data. The prime objectives of this proposed research is to provide a hybrid algorithm for applying on smart data. And feasibility of running machine learning algorithms on Big Data Frameworks and optimization of algorithms for big data [9].

The prime focus of the proposed research is to provide insight how data sets are handled by learning algorithms. The quantification of results and learning pattern of various algorithms provides a source to make to future decisions. In addition, the research provides understanding of machine learning algorithms for comprehension of smart data sets generated by IoT. At the very first instance the smart data is generated by IoT devices through inbuilt applications. This smart data generate with specific domain can be fetch to a machine learning model for resolving the issues arising out of the growing use of devices and data. In case of real time applications, there should be consideration of response time and reliability. This can be achieved only if a learning algorithm would be used properly with prior compatibility with the speed of data creation. The accuracy level should also be a prime criteria for handling the data. Because the accuracy can only provide us better results after stage of data execution. The machine learning algorithms reveal may insights regarding the data characteristics. To explore the more insight into the smart data, the data patterns must be looked into detail. This pattern finding and extraction of data will help in enhanced accuracy score, event will be responded on real time basis and it will consequently affect the decision making. The Random Forest has less training time and multiple trees minimizes risk of overfitting. Moreover, this machine learning algorithm performs better on big data, for data sets with large size, highly accurate predictions are produced. In addition Random Forests can maintain accuracy when a large portion of data is missing.

Random Forest or Random Decision Forest is method that operates by constructing multiple Decision trees during training phase. The Decision of the majority of the trees is chosen by the random forest as final solution.

The other details of this proposed research follows as. Section 2 discusses the related work in the area of handling data sets. The section 3 provides the basic concept of smart data technologies. The proposed hybrid algorithm is discussed in Section 4. The results and observation from experimental work can be viewed in Section 5. The research is concluded with future scope in Section 6.

II. Literature Review

Oscar D. Lara et al. surveyed on wearable sensors for the human activity recognition in the state-of-the-art domain. The parameters used where learning scheme and response time for introducing the organization of human activity recognition systems in two-level-taxonomy format. The scheme qualitatively compared 28 systems with regard to parameters viz general design issues, response time, flexibility, obtrusiveness, recognition accuracy and other issues. The important parts like machine learning and extraction is included as components of human activity recognition systems. The authors have provided further directions to explore in more pervasive and realistic scenarios [10].

The researchers have used smart phones to present an efficient way to classify the activity of humans on daily basis. The basic design has been simulated by considering the fixed-point arithmetic of Support Vector Machines methodology. The research has using principles of Structural Risk Minimization in which complex approaches are neglected instead simpler techniques are used which provides equivalent attributes to learn. The use of the proposed research is to update the ambient applications using current technology such as in smart and remote patient monitoring environments. The properties like minimal use of resources and real time processing which saves the energy overhead for maintaining recognition is providing advantage over traditional approaches [11].

A novel approach has been proposed by the authors based on sensor worn on body [12]. A two-phase algorithm with abnormality detection has been proposed for looking into abnormal activities when there is scarcity of training data. SVM one-class in phase first is built of normal activities for filtering the normal instances of the activities. The adapted abnormal activity models are used to track suspicious traces using KNLR. The researchers claimed that proposed approach provides a better results and tradeoff between false alarm rate and detection rate. The effectiveness of the approach is demonstrated for real data obtained from human body using sensors. The authors have also described drawbacks of their work that if abnormal activities will become normal, then there is risk in general abnormal models. The type of situation is arise when a user repeats the things alternatively after fixed instances of time [12].

A detailed aspects of human activity recognition for offline and real time processing has been presented by Bishoysefen et al. [1].The researcher explained best features of classification algorithm for achieving realtime optimization for recognition accuracy and computational complexity. The data from more than 10 sources has been collected based on day to day activities and exercises. The result of analysis showed that machine learning algorithms have better performance with regard to the efficiency and accuracy.

The big data is amazing source of supply of useful knowledge and information for different end-users and varied systems. In order to handle such kind of inflow of information, the automation is a reliable source and that can be achieved using processing through machine learning. The information and communication technology is serving in various analysis sectors by providing specified tools and platforms for making professionals enable get valuable predictions. These ICT based techniques are developed by prominent firms viz. IBM, Microsoft, Google, etc. The research published provides concepts of Machine learning algorithms in Big Data Analytics [13].

The authors researched the work done on Internet of Things using the big data mining concept. The researchers provided three tiers for stage wise analysis of data in future [14].

The researchers [15] designed a random forest algorithm with improved classification for multiple classes related to a disease. The algorithm provides better classification of individual variable. The improved method has increased accuracy and general process of work. The percentage increase for the classification accuracy was achieved upto 97.80% for multi-class dataset.

The data missing in data sets is creating large number of classification errors. There is a technique called imputation technique that helps to complete data which is having missing datasets. The researcher [16] developed an approach for incorporation of feature selection of genetic-based method and imputation for enhancement of classification of missing or incomplete data.

The internet among various objects is called (IoT) using sophisticated and complex communication technology without human intervention. The researchers put forth a big data based analytical healthcare system using Random Forest algorithm. The methodology proposed shows better accuracy for classification than traditional logistic regression and Gaussain method [17].

The research presented in [18] focuses on smart network fault analysis prediction. The proposed research used a modified RF algorithm for providing enhancement in accurate analysis prediction. The methodology proposed improves accuracy of overall system.

The study provides details about how machine learning techniques can become base for smart data analysis for IoT. The deployed methodology provides high velocity data processing. In addition fast training and classification of datasets have been achieved [19].

The presence of noise during classification causes incorrect labeling in data. The situation becomes disastrous as it changes the basic variables and instances. The researchers [20] put forth an ensemble method which is iterative in nature for restricting noisy instances. The proposed methodology effectively contributes in transforming simple big data to smart one.

The authors [21] proposed a novel approach to deal with big data in which Random Forest algorithm and Support Vector Machine is used. The feasibility and robustness check is performed using parameters like confusion matrix, recall, precision, specificity and sensitivity. The result shows 95% accuracy with big data.

The researchers [22] proposed a big data framework with scalability which collects data from smart devices and stores that in NoSQL. The machine learning algorithms has been imbibed with framework for future predictions. Various machine learning algorithms have been used to load forecast domain specific environment.

The research in [23] focuses on making smart data out of large chunks of raw data. The researchers have used two big data libraries BigDaPSpart and BigDaPFlink for extracting smart data from big data. These libraries are built on Apache Spark and Apache Flink big data frameworks for cleaning, discretization and other things.

The authors in [24] have integrated smart persistence algorithm, past production data, irradiance along with Random Forest machine algorithm for enhancing capability of producing forecasts accurately. The methodology has shown incredible results.

The researchers have proposed a framework for extracting hidden patters with data stored in database. The research was carried out for looking into future possibility of treating vulnerable diseases like breast cancer, diabetes, heart diseases etc. Machine learning algorithms such as Random Forest and others have been used for prediction analysis. The results showed that Random Forests provide better and accurate results than others [25].

III. Smart Data Technologies

The growing demands of control and co-ordination among various sectors and programs has created need to make the novel systems to be smart enough to take care of even a minute detail. The general architecture of smart exigencies is shown in Fig .1. The system has to co-ordinate with various domain specific data stores and provide an optimal solution in the form of analytics reports [26].

Fig. 1 provides a detailed pictorial representation of the smart data technology program. The use of Big Data analytics and techniques to extract patterns and solutions for a domain specific system incorporated with issues related to security, social and economic aspect, legal aspect and many more. In order to deal with problems that depends on large data sets and technology uses follow some constraints that are used in all digital solutions provided. The program associated with smart data generally provides three issues described that should be incorporated in addition to achievement of smart analytics of big data through machine learning.

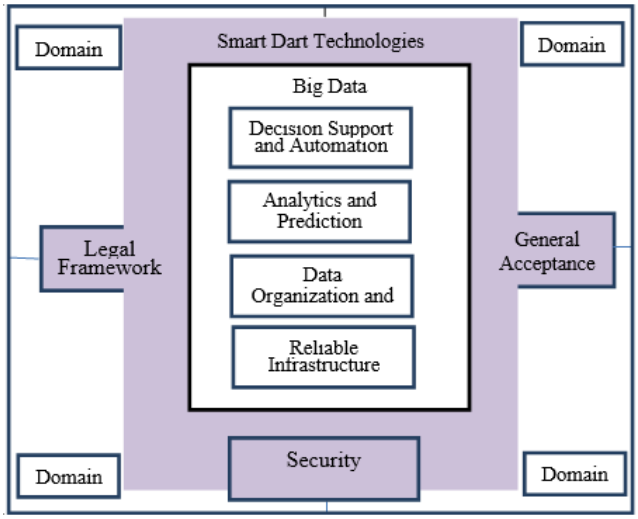

Fig. 1. Smart Data Technology Program [26].

IV. HYBRID ALGORITHM

The research proposed here is based on hybrid algorithm that is used to implement the machine learning at node level. Each node is an analytical unit it is related domain [27]. The node is a mode which provides a storage unit and the related functionality to analyses the data present in node. In current context Random Forest (RF) is an important concept to convert a simple data set to a smart one. This machine learning algorithm is flexible algorithm. The prime functionality of this algorithm is to provide accurate and efficient results, with using any tune of type hyper-parameter [25]. The RF is researcher's prime importance because of its simplicity and mainly to provide definite classification and regression. The Random Forests algorithm comes under the category of supervised learning algorithm. The algorithm creates a random forest decision trees which are trained with "bagging" methodology. The Random Forests provides multiple trees to classify on basis of attributes a new object [26]. Bootstrapping the data plus using the aggregate to make a decision is called "Bagging". The general flow of algorithm is shown in Fig. 2 reference node architecture. The data is collected from reliable source. The source can either be online source for current data or it can be archival data from any storage medium. The data is fetched on basis of some query to be supplied for analytics purpose. The algorithm proposed is integration of similarity formula with Random Forest Machine Learning algorithm.

The section describes the general working of the proposed algorithm that is based on machine learning concept random forests. The trees constructed using this ML algorithm works independently. A rectangular matrix is used to represent nodes of a tree which in turn represent the data stored in each domain node, and at each step of the construction the cells associated with leafs of the tree form a partition of matrix. The root of the tree corresponds to all of matrix.

At each step of the construction a leaf of the tree is selected for expansion. In each tree we partition the data set randomly into two parts, each of which plays a different role in the tree construction. We refer to points assigned to the different parts as structure and estimation points respectively. The shape of a given tree is influenced by allowing structure points. The internal node of the tree is determined by the split points and dimensions. The predictions made by the leafs of the tree are not by any way intervened by the structure points. There is a dual role played by the estimation points. There is no effect by estimation points on shape of partition of trees but the estimation points fit the values of estimators in each leaf. The assignment of points to estimations or structures with original equivalent probability has data that is randomly distributed among each trees. The partitions ensures the real time consistency in the trees formed out of data. There is no need to add parts for fitting each and every subset of data with the tree because making more samples lowers the performance of the system. The construction of tree is kind of parameterized which provides minimum estimations that should appear in a leaf. The size of the training set is taken into consideration while setting the parameters. The training set size also provides information about the minimum size of leaf [28].

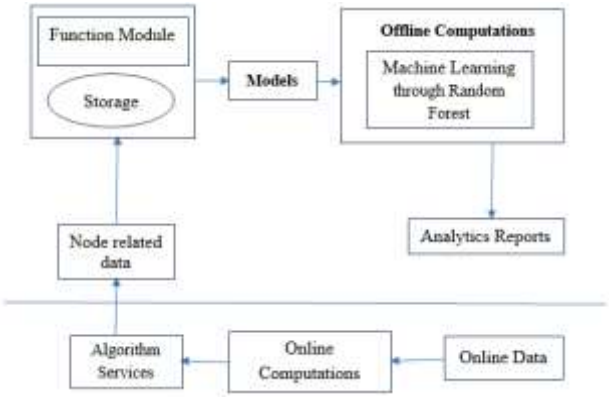

Fig. 2. Reference Node Architecture.

There is random selection of minimum of (1+Poisson((λ), D) on basis of distinct candidate dimensions when there is selection of leaf for expansion. The candidate split points are traversed for choosing split point in a leaf in case of each dimensions of candidate. In case of standard algorithm of random forest, the splits points are in a specified range and the candidate dimensions is projected into these points. At time of searching process, the range is declared by selecting specified points and process of search is done only over the defined set of points. However, the hybrid algorithm changed the traditional process by restricting the range to the specified set of structure points. And restring rage enables the trees to be in balanced state [28].

$$Error(L) = \frac{1}{N^s(L)} \sum_{\substack{Y_j=A \\ I_j=S}} (Y_{j\_} \bar{Y}^L)^2 \quad (1)$$

$$I(S) = Error(L) - Error(L') - Error(L'')$$

The L represents the leaf to split and L′,L″ represents the children which splits L at S. The empirical mean is denoted by notation $\bar{Y}^L$*for* structure points falling in the cell *L*and the number of structure point counts in L is denoted by.

Ns (*L*). whether point (X_j, Y_j) is an estimation point or structure is denoted by indicator variables $Ij \in \{e, s\}$. When *I(S)* maximizes without creating any further children with less than k_n estimation points, the split is chosen as candidate and for non-existence of any such candidate stops the expansion [28].

$$f_n^i(x) = \frac{1}{N^e(A_n(x))} \sum_{\substack{Y_j \in A_n(x) \\ I_j=e}} Y_j \quad (2)$$

The predictions of each tree are averaged by:

$$f_n^{(P)}(x) = \frac{1}{P} \sum_{j=1}^{P} f_n^j(x) \quad (3)$$

The various similarities has been implemented and an integrated method has been put forth for minimization of errors. The similarities has been defined through random mathematical functions. The Jaccard index has been used when the data is of discrete nature to check the errors [29]. The formulae specification are:

$$\text{Jaccard Similarity } (j_i j_p) = \frac{|J_i \cap J_p|}{|J_i \cup J_p|} \quad (4)$$

Where

$Ji= \{u \in U | r_{u,j} > 0\}$ *and* $J_p = \{u \in U | r_{u,p} > 0\}$.

Triangle Similarity

$$(j_i,j_p)= 1 - \frac{\sqrt{\sum_{u \in C_{j_i,j_p}} (r_{u,i} - r_{u,p})^2}}{\sqrt{\sum_{u \in C_{j_i,j_p}} r_{u,i}^2} + \sqrt{\sum_{u \in C_{j_i,j_p}} r_{u,p}^2}} \quad (5)$$

[0,1] is the range value, and 0 indicates $C_{ji} = \emptyset$ [4]

A. Algorithm Description

Algorithm(Nodes, N, S)
Nodes represent the set of nodes in a domain
N is number of Nodes with varied sets of clusters &
a)For each nodes N in Nodes
If the Key matches the node N
For each sub node S
Create a fuzzy random forest using formula(1-3) for different variations
For each decision tree in forest
If the input variable is incessant
For each split point
Create a partition using formula-5;
If there is definite variable
Calculate the similarity using formula-4
b) then choose the value with the optimalindex;
c)accordingly provide child nodes based on the output produced using the indexes;
d) the calculate the gravity of association of each value with the next level nodes;
e) Repeat all the steps a-d until for every node in a tree
end

The general declarations of similarities in the coded form is written as follows:

```
public class Generatecosine {
      public double cosval()
   {
      double cv=Math.random();
      return cv;
   }
    public double jacardcal()
   {
      double jv=Math.random();
      return jv;
   }
}
```

The definitions are generally coded in a format shown below.

```
Generatecosine gn = new Generatecosine();
for(int i=0; i< similaritycount;i++)
{
    if(allcosine[i] ==0)
      {
           allconsine[i]=(float)gn.cosval();
      }            System.out.println(similarityvalue[i][1]+""
  + " "+similarityvalue[i][2]+""
   +"     "+allcosine[i]);
}
```

V. Results Discussion

The proposed algorithm gathers information from various data sources. The data files are distributed into the various domains through predefined algorithm [27]. The further process is done in internal nodes where the actual acquisition, management and mining is performed for analytics process. The new smart data set based algorithm proposed performs the optimal analytics of data by following the concepts of machine learning. The improved machine learning algorithm through incorporation of multiple similarity index is providing the environment for error free analytics in real time manner. The nodes present inside the domain are maintaining the security while loading data for mining purpose [27]. The proposed algorithm enhances the services and gives generalized machine learning Big Data techniques and various different protocols.

The algorithm provides the services in three phases. In phase first, the data mining process is done and the smart data clusters are fetched into the program for making the trees using enhanced Random Forest algorithm. Also in the same phase, the hybrid algorithm has been tested on twelve distinct datasets. The datasets has been obtained from Kaggle online data repository and is tabulated in Table I. The granularity based approach for data stream mining is good method so far as computational intelligence is concerned [30].The datasets contains varied number of input attributes and instances. The varied folded validation has been performed with more than two trials with different basic attributes for the machine learning algorithm. In order to check the exact result the number of folds and validations has been kept same in each dataset in order to check the overall functionality. This phase generally provides the classification accuracy of proposed algorithm. Fig. 3 shows the classification accuracy comparison of the traditional and proposed algorithm and which clearly signifies the outstanding performance of the system in consideration. The red dots are showing the enhanced results so far as classification accuracy is concerned.

TABLE I. Classification Efficiency of Hybrid Algorithm

Disease Type	Instances	Attributes	Classification Efficiency	Classification Efficiency of Hybrid Algorithm
COVID-19	457	10	94.44	98.32
EBOLA	145	8	87.06	95.3
MERS	121	5	78.32	86.15
H1N1	175	10	86.06	96.06
SARS	208	12	72.45	89.35
HIV/AIDS	193	8	92.36	99.12
H3N2 HonKong Flu	113	10	84.16	92.34
H2N2 Asian Flu	241	6	89.02	96.56
Spanish Flu	243	5	78.12	87.63
ZIKA	54	6	79.11	89.6

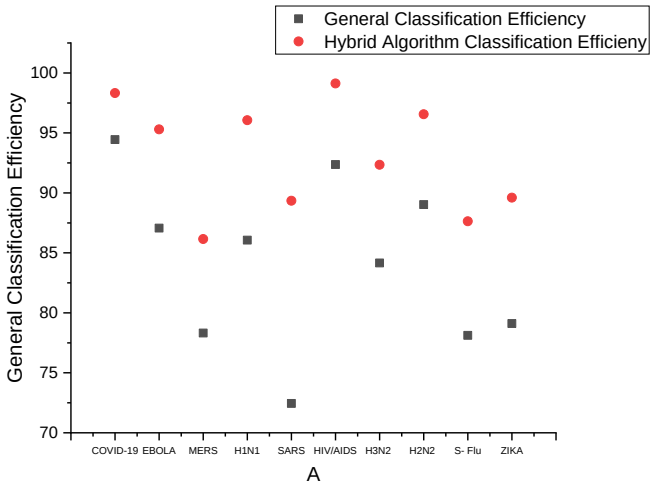

Fig. 3. Classification Accuracy of Hybrid Algorithm.

The phase-II provides verification details of error measures enhancement of proposed algorithm and details of various kinds of error measures. The error specification of results general machine learning algorithm and the proposed algorithm has been compared by incorporation of new rule called integrated method of Traingle, Jaccard similarity. The classical machine algorithm was executed with smart datasets proposed in [28]. Then, the analysis of data present in the leaf nodes was done for root mean square error and absolute error. Fig. 4 shows the performance of all the datasets for different data set when the attributes are selected less in number. The procedure proposed in [31] has been adopted to measure the difference in errors. The least error has been shown by the grey color integrated method. Table II shows the extent of data and its relative measures of error reduction for various similarities.

Table III provides the data related to error reduction when the number of attributes selected have been increased by 10% than earlier in k range. In Fig. 4, there is comparison of Mean Absolute Error using various similarity concept of measures.

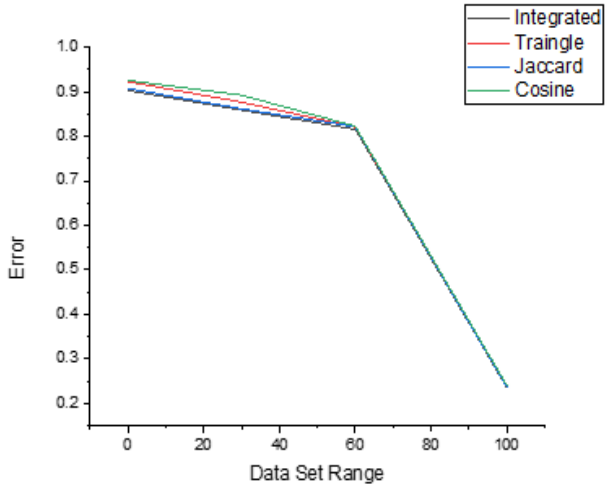

Fig. 4. Mean Absolute Error by Proposed Algorithm using Integrated Formula.

TABLE II. MEAN ABSOLUTE ERROR ON VARIED INPUTS

Measure /Dataset Size	**10 K**	**50 K**	**100K**	**1M**
Cosine	0.732	0.696	0.625	0.187
Jaccard	0.711	0.674	0.617	0.18
Triangle	0.724	0.688	0.621	0.183
Integrated	0.707	0.671	0.614	0.179

TABLE III. ROOT MEAN SQUARE FOR HYBRID INTEGRATED ALGORITHM FOR K DATA (HIGHLIGHTED)

Measure/Dataset	**Flu Data Set 100K**
Cosine	0.748
Jaccard	0.729
Triangle	0.721
Integrated	**0.713**

The general systems cannot complete the algorithm with a measurable period of time whereas, integrated similarity achieves the optimal/best values for mean absolute error. The datasets used showed lowering of the percentage errors by a significant value about 6% in two data sets, 7%, 14% and 24% in other data sets respectively than the results from general methods. The results are shown in Fig. 5 for various similarity measures and the proposed method always performs best among others. The percentage change in error depends on accuracy of input data set.

Table IV provides the data related to error reduction when the number of attributes selected have been increased by 10% than earlier in m range. Fig. 6 shows details about lowering Root Mean Square Error. The reduction of errors as shown by using novel approach is 7.2 %, 17%, 21% and 12% for smart datasets. The traditional similarity measures cannot get completed within measuring time unit whereas, same is accepted by the new integrated measure. The results obtain are optimal/best for the root mean square error. The results clearly shows that optimal machine learning is adapting to the changes made in the input data sets.

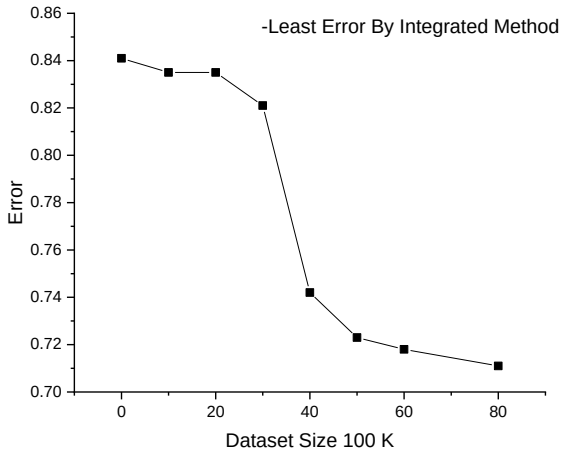

Fig. 5. Root Mean Square Error b for Dataset of Size 100k.

TABLE IV. ROOT MEAN SQUARE FOR HYBRID INTEGRATED ALGORITHM FOR M DATA (HIGHLIGHTED)

Measure/Dataset	**Flu Dataset for 1M**
Cosine	0.681
Jaccard	0.662
Triangle	0.679
Integrated	**0.659**

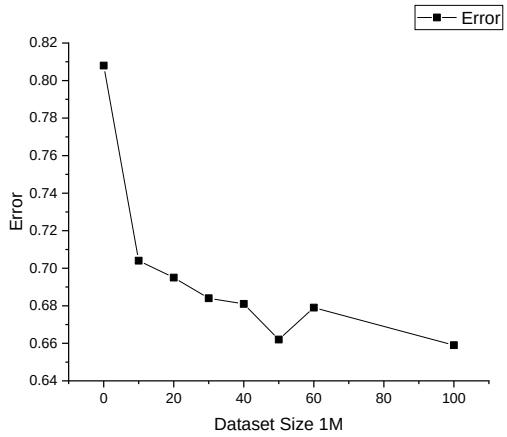

Fig. 6. Root Mean Square Error for Dataset of Size 1M.

VI. CONCLUSION AND FUTURE SCOPE

The inflow of smart applications is having lot of impact on general overall development. The presence of such applications will improve the standard of living. The incorporation of Big Data analytics into difference applications will provide decision making and effective intelligent policies. The research proposed provides a hybrid algorithm for using smart data sets from IoT systems for future trend analysis, general systems co-ordination and accuracy. The proposed algorithm provides a layered model for various operations viz. processing, aggregation, filtering and transmission of big data. The prime role of this proposed algorithm is to make the decision trees classified by machine learning Random Forest concept. The proposed scheme also provides optimal throughput with greater variance over time. To reduce the variation of the throughput, a cross layer model will be considered in future work. The model will look further into heterogeneous challenges.

REFERENCES

[1] Sefen, S. Baumbach, "Human Activity Recognition Using Sensor Data of Smart phones and Smart watches,"/ ICAART 2016.12.

[2] H. Green, "The Internet of Things in the Cognitive Era: Realizing the Future and Full Potential of Connected Devices," Technical Report; IBMWatsonIoT: New York, NY, USA, 2015.1.

[3] J. Gubbi, R. Buyya, S. Marusic, Palaniswami, "M. Internet of Things (IoT): A vision, architectural elements, and future directions. Future,"Gener. Comput. Syst. 2013, 29, 1645–1660.

[4] D. Evans, "The Internet of Things: How the Next Evolution of the Internet is Changing Everything", CISCO White Paper 2011. 1, 1–11.

[5] J. Manyika, M. Chui, P. Bisson, J. Woetzel, R. Dobbs, J.A.D. Bughin, , "Unlocking the Potential of the Internet of Things; Technical Report; McKinsey Global Institute: New York, NY, USA, 2015.4.

[6] I. Lee, K. Lee, "The Internet of Things (IoT): Applications, investments, and challenges for enterprises," Bus. Horiz. 2015, 58, 431–440.

[7] Y. Yoo, O.Henfridsson, K. Lyytinen, "The New Organizing Logic of Digital Innovation: An Agenda for Information Systems Research," Inf. Syst. Res. 2010, 21, 724–735.

[8] F. Wortmann, K. Fluchter,"Internet of Things: Technology and Value Added," Bus. Inf. Syst. Eng. 2015.57, 221–224.

[9] M. T. Yazici, S Basurra, M. MGaber, "Edge Machine Learning: Enabling Smart Internet of Things Applications," Big Data Cognitive Computing. 2018, 2, 26; doi:10.3390/bdcc2030026.

[10] O. D. Lara and M. A. Labrador, "A Survey on Human Activity Recognition using Wearable Sensors,"/ IEEE 2013.

[11] D. Anguita, "Energy Efficient Smartphone-Based Activity Recognition using Fixed-Point Arithmetic,"/ JUCS 2013.10.

[12] J Yin, "Sensor-Based Abnormal Human-Activity Detection,"/ IEEE 2008.11.

[13] K. Sree Divya, P. Bhargavi, S. Jyothi, "Machine Learning Algorithms in Big data Analytics," International Journal of Computer Sciences and Engineering, Vol.6, Issue.1, pp.63-70, 2018.

[14] S. Shadroo, A.M. Rahmani, "Systematic survey of big data and data mining in the internet of things," (2018) Comput Netw 139:19–47.

[15] A. Paul, S. Rho, "A probabilistic model for M2M in IoT networking and communication," (2016) Telecommun Syst 62(1):59–66.

[16] C.T Tran, M. Zhang , P. Andreae, B. Xue , L.T. Bui, "An effctive and effient approach to classifiation with incomplete data," (2018) Knowl Based Syst 154:1–16.

[17] K. Lakshmanaprabu, K. Shankar, M. Ilayaraja, N. A. Wahid, V. Vijayakumar and N. Chilamkurti, " Random forest for big data classification in the internet of things using optimal features," (2019) International Journal of Machine Learning and Cybernetics. 10. 10.1007/s13042-018-00916-z.

[18] R. Lin, Z. Pei, Z. Ye, B. Wu, G. Yang, "A voted based random forests algorithm for smart grid distribution network faults prediction", (2019) Enterprise Information Systems. 14. 1-19. 10.1080/17517575. 2019.1600724.

[19] M. H. Alsharif, A. H. Kelechi, K. Yahya, S.A. Chaudhry, ". Machine Learning Algorithms for Smart Data Analysis in Internet of Things Environment," (2020) Taxonomies and Research Trends. Symmetry. 12. 88. 10.3390/sym12010088.

[20] D. G. Gil, F. L.Sánchez, J. Luengo S. García and F. Herrera, "From Big to Smart Data: Iterative ensemble filter for noise filtering in Big Data classification," (2019) International Journal of Intelligent Systems. 34. 10.1002/int.22193.

[21] B. Devi, S. Kumar, Anuradha and V.G. Shankar, " AnaData: A Novel Approach for Data Analytics Using Random Forest Tree and SVM," (2019) In: Iyer B., Nalbalwar S., Pathak N. (eds) Computing, Communication and Signal Processing. Advances in Intelligent Systems and Computing, vol 810. Springer, Singapore. https://doi.org/10. 1007/978-981-13-1513-8_53.

[22] S. Oprea and A. Bâra, "Machine Learning Algorithms for Short-Term Load Forecast in Residential Buildings Using Smart Meters, Sensors and Big Data Solutions," in IEEE Access, vol. 7, pp. 177874-177889, 2019, doi: 10.1109/ACCESS.2019.2958383.

[23] D. G. Gil,A.A Barros, J.Luengo, S. García and F. Herrera, "Big Data Preprocessing as the Bridge between Big Data and Smart Data: BigDaPSpark and BigDaPFlink Libraries. 324-331. 10.5220 /0007738503 240331.

[24] J. H. Tato, "Using Smart Persistence and Random Forests to Predict Photovoltaic Energy Production," (2018) Energies. 12. 100. 10.3390/en12010100.

[25] P. Kaur, R. Kumar and M. Kumar, " A healthcare monitoring system using random forest and internet of things (IoT). Multimed Tools Appl 78, 19905–19916 (2019). https://doi.org/10 .1007 /s 11042-019-7327-8.

[26] BMWi, "Smart-data-technologien (German)," Smart Data Accompanying Research, Tech. Rep., 2015. [Online]. Available: https://www.digitale-technologien.de/DT/Redaktion/DE /Downloads /Publikation/smartdata_brochure_english.pdf?__blob=publicationFile&v =18.

[27] D. Masroof, Munishwar Rai, "A Novel Framework for Enhancing QoS of Big Data", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 4, 2020.

[28] M. Denil, D. Matheson and N. D. Freitas, "Narrowing the Gap: Random ForestsIn Theory and In Practice", Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

[29] P. Jaccard, "Nouvelles researches sur la distribution florale," Bull Soc Vaud Sci Nat. 1908; 44:223±270.

[30] M. M. Gaber, "Data stream mining using granularity-based approach. In Foundations of Computational, Intelligence," Volume 6; Springer: Berlin/Heidelberg, Germany, 2009; pp. 47–66. 8.

[31] S-B. Sun, Z-H. Zhang, X-L Dong, H-R Zhang, T-J. Li, L. Zhang, F. Min, "Integrating Triangle and Jaccard similarities for recommendation," (2017) PLoS ONE 12(8): e0183570. https://doi.org /10.1371/journal.pone.0183570.