DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4164

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

ROBUST FUZZY NEURO SYSTEM FOR BIG DATA ANALYTICS

Aman Taneja
M.Tech ,Computer Science & Engineering
UIET, MDU
Rohtak, India

Mrs. Kamna Solanki
Assistant Professor (CSE)
UIET, MDU
Rohtak, India

Dr. Sandeep Dalal
Assistant Professor
DCSA, MDU
Rohtak, India

Abstract: Big Data is the relationship of data size and its processing speed. These are so big and complex that traditional data processing software is inadequate. These days it's a high challenge to construct architecture to take out information cost-effectively from huge, volume of data at remarkable rate. So, there is a need to find economical and valuable solutions for the major challenges of fatly growing volume and concern. Through this paper, we can become proficient in big data analytics, its tools and application areas. It also presents uncertainty issues related to Big Data for which the solution we provided by combining fuzzy and neural network concepts to assemble a new intelligent system ANFIS that has brought together tendency to get the results by relating knowledge representation, uncertainty and modelling the key feature of big data to provide a superlative solution.

Keywords: ANFIS, Fuzzy System, Membership Function, Neural System, Uncertainty

I. INTRODUCTION

Over the last two decades we can find repository of data to be generated digitally on the web, social networking sites, from mobile devices or online transactions. Data is complex and making it useful information is the main concern to make it acceptable and appreciated by users or learners. Prepare the huge data sets to make meaning out of it is the main task in big data. These enormous amounts of data sets are increasing with passage of time at a great vast scale that are generally petabytes and zettabytes to exabytes thus known as Big Data.
The 5V's essential characteristics of Big data i.e. volume, velocity, variety, value and veracity.
Data is developed from various sources in a massive amount and flowing progressively generating some appropriate business insights in a protected aspect.

This massive volume is categorized into 3 categories specifically

- Structured data having SQL databases that can be text or integer values that are specific to them.

- Unstructured data cannot be fixed into relational database schemas. Webpages, PDF files, PowerPoint presentations, emails, document files are some of the examples that fall into this class.

- Semi structured data constitute of both the structured as well as unstructured data. Word processing software, NoSQL databases, weblogs and social media feeds.

It is difficult to reserve and consider huge datasets using normal software tools. Now these days' new technology are possible to take care of this Big Data. Big Data administration is now a challenging technology for new generation.

Data is often shared or separated and replicas are made on each system to avoid breakdown but it increases duplicity. It was Google who recognized the significance of Big data. The aim of this program is to deposit the significant report from huge load data sets effortlessly and removing storage, fault tolerance and scalability issues and thus guiding it to benefit in the market for profit making.

Encapsulating diverse and composite data from different range and securing the condition of data for operating compulsory task by machines is an important factor. Data is burst in a extensive form, this huge density data needs to be figure out and put into use.

II. BIG DATA ANALYTICS

Big data inquiry is defined to be a channel of
Acquisition, extraction, cleaning, integration, aggregation and visualization, analysis and modeling; and interpretation as shown in Figure 1

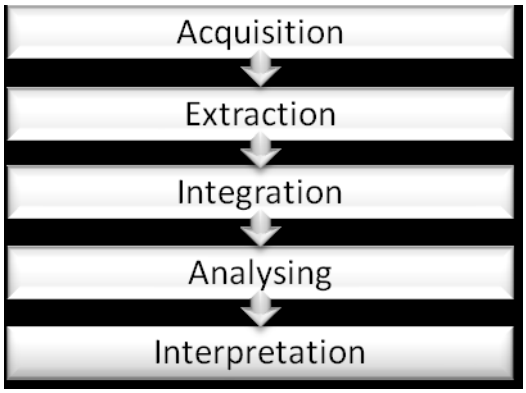


Figure 1 Phases of big data analysis [13]

Previously, all the effort in big data analytics was based on negotiable database which has been switched to unstructured or semi structured datasets based on data mining and machine learning techniques.

III. APPLICATION AREAS

There are many assumption of Big data in day to day life that has made changes and discovery grown in this real world. It is meant on researchers, patients, business tycoons, government, and stakeholders.
In Health department, there is enormous amount of data stored regarding patient's medical data including test reports, prescriptions, follow ups that are recorded by the medical practitioners. Correlating the medical history of patients with the drug manufacturing companies is complex. In Security, activities related to criminal activities can be predicted using big data analysis by detecting fraud transactions in banks and to halt terrorists plans by various security agencies. Net surfing, online shopping, making social networks on web is a big source for big data analytics. It provides maximum utilization of data on the web these days and thus creating issues for managing it. By using real time traffic information, there will be ease of traffic management and optimized route can be profound using advanced analytics. Anything that includes transaction in result of various operations conducted on web takes the picture of trading. Along with that, taking log of all the transactions and maintaining it requires analytics. There are various big data algorithms driven for trading for the benefit of financial traders [15]

IV. TOOLS

To cope with the situation various technologies are introduced namely Hadoop, Map Reduce which again follows distributed file system.

A. Hadoop

An open source Java framework technology and library proposed by Apache Software Foundation was created by Goug Cutting[8] and Mike Cafarella in 2005 which act as distributed search Engine Project named as Hadoop (Highly Archived Distributed Object Oriented Programming).[3] The technology is built with the aim of processing large data sets with several servers to work in a distributed manner and achieve benefit of cost, time and storage efficiency. It is a tool that can handle hundreds and thousands of computers with fault tolerance and scalability detection. [1]

A single cluster of Hadoop contains one Master node and multiple slave nodes.[9] The master node consists of Data node, Name node, Job Tracker and Task Tracker where slave node acts as both a TaskTracker and Data node, each having their tasks for handling structured as well as unstructured data.

Hadoop distributed file system is a file system which is highly fault tolerant and works on low cost hardware. The size of node ranges from 64 Megabytes to Gigabytes values. With high bandwidth clustered storage architecture reduces data loss. But it has the risk of data access, theft as the data is replicated on several nodes, so security aspects increase to protect it from breaches. [3]

B. Map Reduce

A simple parallel programming model for computation on large clusters for substantial scalability of thousands of servers and processing huge data sets. The working is signified by its name "MapReduce" where first mapping is done and then reducing the data sets. The basic idea is to divide the clusters into sub clusters and after applying MapReduce combining them to get the results. It has master /slave architecture [12] with one master node and several slave nodes managed by it. Each node of cluster is broken down into key/value pairs. [4] The different phases included in it are

- sorting,
- partitioning and
- combining values

Different ways have been implemented by Google for possessing work in the field of Big data. Usage of thousand bunch of machines for scalability has not work as it was supposed due to malfunctioning of machines[6][10][11]
Programs are robotically parallelized due to the abstraction designed for simplifying the untidy representation of the computation because of the huge complications faced by the system [7]
So, the idea of key/value pairs was anticipated which worked well for even large amount of data.

In Map
Two defined tuples act as input which generates intermediate values by emitting each value individually after that acts as input for reduce stage

Map(key,value)→ list(key,Intermediate(values))
Emit Intermediate(value,"1")[11]

In Reduce
All the values which are output of map stage in list form are processed in parallel and combined values are generated

List (key,Intermediate(value))→list(values)

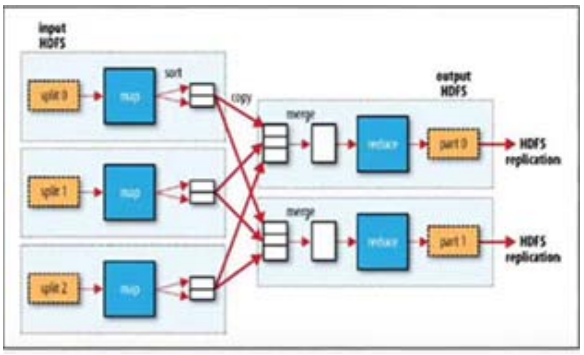


Figure 2: Map reduce data flow with multiple reducer

V. UNCERTAINTY AND HETEROGENEITY OF BIG DATA

Uncertainty is the term for inappropriate and abnormal data which may lead to incorrect
classification of system . There are various factors involved for its existence and in a variety of forms as of dimension errors, noise, processing issues or faulty data management.

In big data, uncertainty can be caused due to large volume of data, miscellaneous and ever changing sources of data, unstructured and different data formats [20].Anything below 100% is uncertain for scientists and researchers so it is vital to toil uncertainty for decision making.
Veracity plays an important role in dealing with uncertainty aspect. Monte-Carlo sampling is one of the ways to deal with uncertainty in spatial.
Different ways to tackle uncertainty are to construct optimized algorithms and use advanced mathematics for calculation of probabilistic distributions [21]

To solve complex machine learning problem a technique of combining Fuzzy System Neural network for big data came into existence named as ANFIS. This collective intelligence system plays crucial role by combining feature of both system and removes some of the limitations of each other by using few artificial intelligence technique and algorithm which makes system performance very well.

Table 1. Comparison between Neuro and fuzzy logic

	Fuzzy System	Neural Network
Interpretable	Yes	No
Fault Tolerance	No	Yes
Knowledge depiction	Yes	No
Learning capability	No	Yes
Explanation capability	Yes	No

In Table 1 it is understandable that Neural network have the learning capability, fault tolerance and uncertainty tolerance while Fuzzy inference systems are very apt for information representation, interpretability as well as explanation and analysis of data.. But they both gives good results in uncertainty tolerance, adaptability and imprecision tolerance.
It is completely human network with set of rules and contain multiple layers. Its motive is to borrow learning capability from neural network and put it into fuzzy system, that is, superimposition of fuzzy over neural network so the fuzzy inference system behaves similar to neural network and thus good modeling and computation is achieved. The ANFIS architecture is shown in figure 3

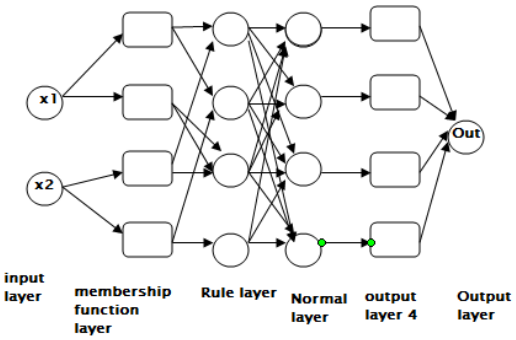


Figure 3. Architecture of ANFIS

VI. RESULTS

Expert knowledge and fuzzy systems to represent big data , neural network for making if and then rules and change input/output membership function to progress the overall performance of the system.[18] to analyze output by ANFIS(Adaptive Neuro Fuzzy Inference System) in MATLAB by using the datasets in which 3 inputs are applied and one output is produced where ANFIS perform as a classifier with inputs Uncrtn , Krep ,Mod respectively.

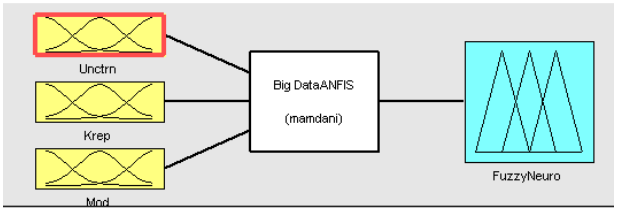


Figure 4. Generating Fuzzy Neuro Output

Here, Uncertainty and Modeling are provided with values Yes if the value is Taken Else No. And Knowledge representation has three levels designated by Poor, Good and Excellent.

Working

The data sets are applied with rules to generate Neuro Fuzzy output that will give result 1, only if Uncertainty and Modeling gives input as Yes.

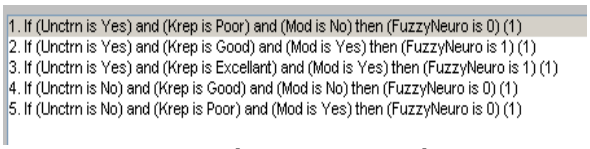
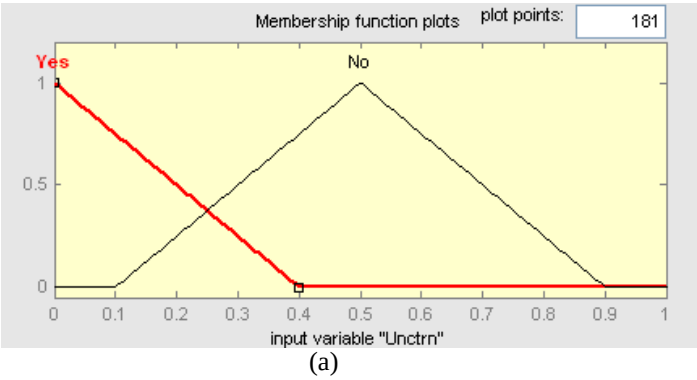
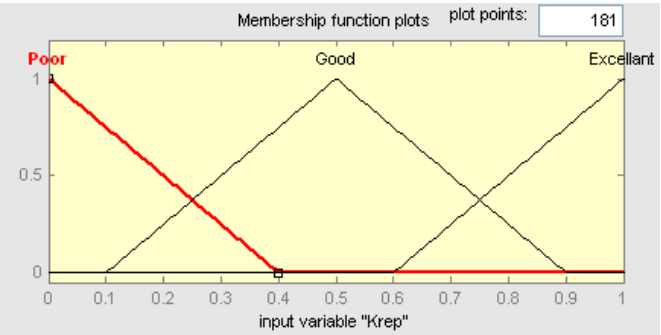


Figure 5. Using dataset to prepare rules

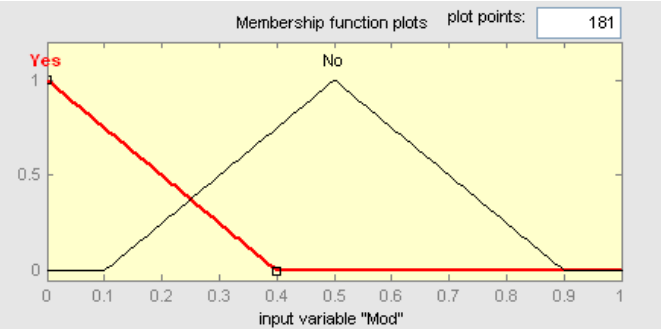
Membership function plots are produced from the training data set as an important role in the network. An inappropriate choice of membership functions can lead to erroneous descriptions.[19] The learning algorithm works in offline mode to check the error rate. We have used a supervised learning system that is Mamdani Neuro-fuzzy system. The Membership function plots are shown in figure 6. Here, Membership functions are plotted w.r.t each input and its corresponding output



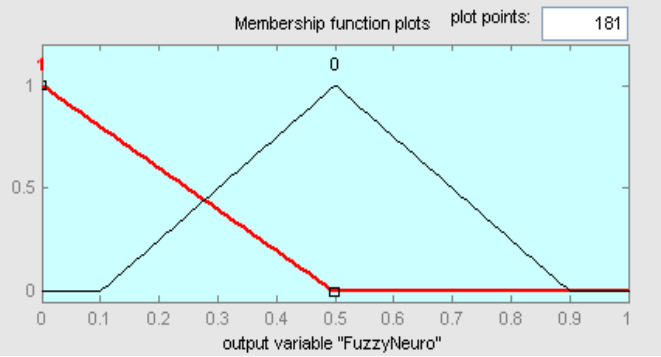
(a)



(b)



(c)



(d)

Figure 6 Membership function plots (a) Uncertainty (b) Knowledge Representation (c) Modeling (d) Fuzzy Neuro

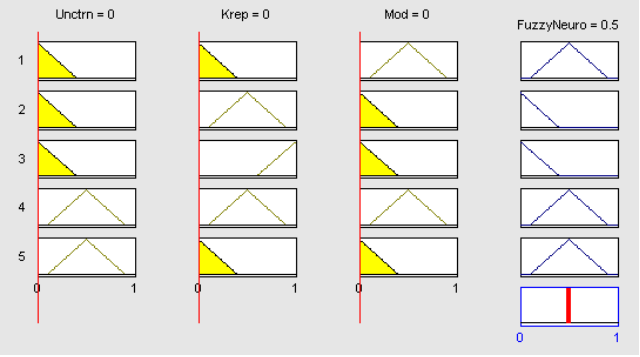


Figure 7 Rule Viewing with inputs [0 0 0]

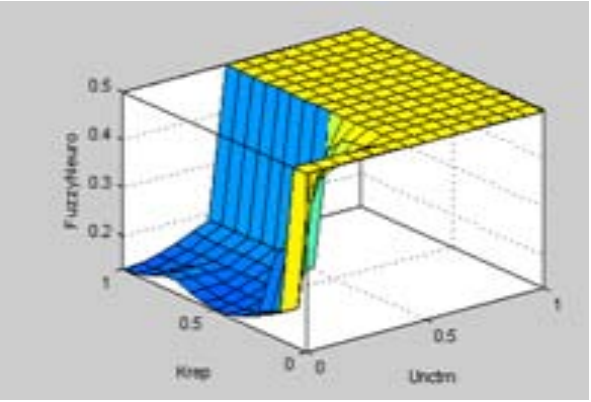


Figure 8 Surface Viewing of Adaptive Fuzzy Neuro System

Figure 8 shows the surface view of the output generated by applying Fuzzy Neuro logic on Uncertainty, Knowledge Representation and Modeling.

VII. CONCLUSION

Fuzzy Inference system is a framework based on if-then rules and fuzzy reasoning. Neural network is based on feed forward and approximation method. The combination of fuzzy inference system and soft computing techniques like neural network provides a good solution for solving big data problems.

The experimental results show that fuzzy rules are implemented on training set to eradicate uncertainty and resolve complex machine learning algorithm. Keeping this thing in mind we have provided an adaptive supervised learning Mamdani type of fuzzy inference system as a solution in this paper by taking the benefit of learning capability with the power of fault tolerance and explanation competence. Membership function plots generated from training data sets plays an important role to represent different parameters and generate its surface view. The expert knowledge In future, we can think about constructing high performance and accurate computational model of Sugeno-type for big data and we can also compare our system with the proposed model.

VIII. REFERENCES

1) M. Vijayalakshmi, " Big Data Analytics Frameworks Parth Chandarana" , International Conference on Circuits, Systems, Communication and Information Technology Applications,(CSCITA),DOI: 10.1109/CSCITA.2014.6839299
2) Amir Gandomi, Murtaza Haider,Ted Rogers, “Beyond the hype: Big data concepts, methods, and analytics”, International Journal of Information Management, 0268-4012/© 2014.
3) B. Saraladevia, N. Pazhanirajaa, P. Victer Paula, M.S. Saleem Bashab, P. Dhavachelvanc, “Big Data and Hadoop-A Study in Security Perspective”,2nd International Symposium on Big Data and Cloud Computing (ISBCC’15)
4) Xue Qin,Brian Kelley,Mahdy Saedy, “A Fast Map-Reduce Algorithm for Burst Errors in Big Data Cloud Storage”, 10th System of Systems Engineering Conference (SoSE) 2015 978-1-4799-7611-9/15/$31.00 ©2015 IEEE
5) .Fazal-e-Amin, Abdullah S. Alghamdi, Iftikhar Ahmad, Tazar Hussain,” Big Data for C4I Systems: Goals, Applications, Challenges and Tools” Fifth international conference on

Innovative Computing Technology, 978-1-4673-7551-1/ 15/$31.00© 2015 IEEE
6) . U. Selvi, Dr. S. Pushpa, " A Review of Big Data and Anonymization Algorithms", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 17 (2015)
7) J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," Commun. ACM, vol. 51, pp. 107–113, January 2008
8) "Apache hadoop." http://en.wikipedia.org/wiki/Apache Hadoop.
9) "Yahoo hadoop tutorial." http://public.yahoo.com/gogate/
10) hadoop-tutorial/starttutorial.html.
11) S. Manoharan, "Effect of task duplication on the assignment of dependency graphs," Parallel Comput., vol. 27, pp. 257–268, February 2001.
12) Dr. Siddaraju1, Sowmya C L2, Rashmi K3, Rahul M, "Efficient Analysis of Big Data Using Map Reduce Framework ", International Journal of Recent Development in Engineering and Technology ISSN 2347-6435(Online) Volume 2, Issue 6, June 2014
13) Hadoop Distributed File System (HDFS), http://hortonworks.com/hadoop/
14) B.Tulasi, " Significance of Big Data and Analytics in Higher Education", International Journal of Computer Applications (0975 – 8887) Volume 68– No.14, April 2013
15) Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The Google _le system. In 19th Symposium on Operating Systems Principles, pages 29.43, Lake George,
16) New York, 2003.
17) S.Sangeetha, A.K Sreeja, "Science No Humans, No New Technologies No changes "Big Data a Great Revolution"", S.Sangeetha et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 3269-3274
18) Big Data for C4I Systems: Goals, Applications,Challenges and Tools Fazal-e-Amin, Abdullah S. Alghamdi, Iftikhar Ahmad, Tazar Hussain, Fifth international conference on Innovative Computing Technology (INTECH 2015) 978-1-4673-7551-1/15/$31.00© 2015 IEEE
19) Hajar Mousanif,Hasna Sabah, Yasmina Douiji, Younes Oulad Sayad OSER, "From Big Data to Big Projects: a Step-by-step Roadmap", 2014 International Conference on Future Internet of Things and Cloud 978-1-4799-4357-9/14 $31.00 © 2014 IEEE DOI 10.1109/FiCloud.2014.66
20) .http://in.mathworks.com/matlabcentral/fileexchange/29043-neuro-fuzzy-classifier
21) F. Barouni, *, B. Moulin, "An Intelligent Atial Proximity System Using Neurofuzzy Classifiers And Contextual Information "
22) The International Archives Of The Photogrammetry, Remote Sensing And Spatial Information Sciences, Volume Xl-2, 2014 Isprs Technical Commission Ii Symposium, 6 – 8 October 2014

a) Abraham ,Adaptation of Fuzzy Inference System

23) Using Neural Learning,Chapter 3 http://ajith.softcomputing.net
24) Shubhangi G. Khadse, "A Survey of Data Uncertainty in Face Recognition ",International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7623-7625

