ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 12, December 2017

Classifying and Analysis of Big Data using Neural Fuzzy System

[1] G. Sindhu
[1]Assistant Professor, Holycross Engineering College, Tuticorin

***Abstract:*-- In this paper discuss about classifying and analysis of Big Data using Neural Fuzzy Systems. A Neuro-fuzzy system is a fuzzy system that uses a learning algorithm derived from or inspired by neural network theory to determine its parameters (fuzzy sets and fuzzy rules) by processing data samples. A Neuro-fuzzy system can be viewed as a 3-layer feed forward neural network. The first layer represents input variables, the middle (hidden) layer represents fuzzy rules and the third layer represents output variables. The input variables are Big Data (term for Data sets). The middle or hidden layer is used to generate an automatic rule for structured and unstructured data by learning algorithm. In third layer it generates an output. Finally analyse latency, throughput, and fault rate.**

Keywords – Big Data, Neural Fuzzy System, Data Set, Classifier.

1. INTRODUCTION

Big data is data sets that are so voluminous and complex that traditional data processing application software are inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy. 'Big Data' is also a data but with a huge size. 'Big Data' is a term used to describe collection of data that is huge in size and yet growing exponentially with time. In short, such a data is so large and complex that none of the traditional data management tools can store it or process it efficiently. Big data' could be found in three forms: Structured, Unstructured, and Semi-structured

II CLASSIFICATION OF BIG-DATA

Structured: Any data that can be stored, accessed, and processed in the form of fixed format is termed as a 'structured' data. Over the period, talent in computer science have achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and deriving value out of it. The sizes are being in the rage of multiple zettabyte. (1021bytes). Data stored in a relational database management system is one example of a 'structured' data.

Unstructured: Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. Typical example of unstructured data is, a heterogeneous data source containing a combination of simple text files, images, videos. This data is in its raw form or unstructured format. Example of Un-Structured Data is output returned by Google Search.

Semi-structured: Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form, but it is not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in XML file.

III BIG DATA ARCHITECTURE

The architecture has multiple layers.
Big data sources layer: Data sources for big data architecture are in a state of disorder. Data can come through from company servers and sensors, or from third-party data providers. The big data environment can ingest data in batch mode or real-time. A few data source examples include enterprise applications like ERP or CRM, MS Office docs, data warehouses and relational database management systems (RDBMS), databases, mobile devices, sensors, social media, and email.

Data massaging and storage layer: This layer receives data from the sources. If necessary, it converts unstructured data to a format that analytic tools can understand and stores the data according to its format. The big data architecture might store structured data in a RDBMS, and unstructured data in a specialized file system like Hadoop Distributed File System (HDFS), or a NoSQL database.

Analysis layer: The analytics layer interacts with stored data to extract business intelligence. Multiple analytics tools operate in the big data environment. Structured data supports mature technologies like sampling, while unstructured data needs more advanced (and newer) specialized analytics toolsets.

Consumption layer: This layer receives analysis results and presents them to the appropriate output layer. Many types of outputs cover human viewers, applications, and business processes.

IV BIG DATA TOOLS FOR DATA ANALYSIS

Data analysis is the process of inspecting, cleaning, transforming, and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision making.

Open Source Data Tool:
Knime:KNIME Analytics Platform is the leading open solution for data-driven innovation, helping we discover the potential hidden in our data, mine for fresh insights, or predict new futures. With more than 1000 modules, hundreds of ready-to-run examples, a comprehensive range of integrated tools, and the widest choice of advanced algorithms available, KNIME Analytics Platform is the perfect toolbox for any data scientist.

OpenRefine:OpenRefine (formerly Google Refine) is a powerful tool for working with messy data: cleaning it, transforming it from one format into another, and extending it with web services and external data. OpenRefine can help we explore large data sets with ease

R-Programming:The R language is widely used among data miners for developing statistical software and data analysis. Ease of use and extensibility has raised R’s popularity substantially in recent years. Besides data mining it provides statistical and graphical techniques, including linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and others.

Orange:Orange is open source data visualization and data analysis for novice and expert, and provides interactive workflows with a large toolbox to create interactive workflows to analyse and visualize data. Orange is packed with different visualizations, from scatter plots, bar charts, trees, to dendrograms, networks and heat maps.

RapidMiner:Much like KNIME, RapidMiner operates through visual programming and is capable of manipulating, analysing, and modelling data. Rapid Miner makes data science teams more productive through an open source platform for data prep, machine learning, and model deployment. Its unified data science platform accelerates the building of complete analytical workflows from data prep to machine learning to model validation to deployment in a single environment, dramatically improving efficiency and shortening the time to value for data science projects.

Pentaho:The platform simplifies preparing and blending any data and includes a spectrum of tools to easily analyse, visualize, explore, report, and predict. Open, embeddable, and extensible, Pentaho is architected to ensure that each member of our team from developers to business users can easily translate data into value.

Talend:Talend is the leading open source integration software provider to data-driven enterprises. Our customers connect anywhere, at any speed. From ground to cloud and batch to streaming, data or application integration, Talend connects at big data scale, 5x faster and at 1/5th the cost.

Weka:Weka, an open source software, is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a data set or called from your own JAVA code. It is also well suited for developing new machine learning schemes, since it was fully implemented in the JAVA programming language, plus supporting several standard data mining tasks.

NodeXL:NodeXL is a data visualization and analysis software of relationships and networks. NodeXL provides exact calculations. It is a free (not the pro one) and open-source network analysis and visualization software. It is one of the best statistical tools for data analysis which includes advanced network metrics, access to social media network data importers, and automation.

Gephi:Gephi is also an open-source network analysis and visualization software package written in Java on the NetBeans platform. Think of the giant friendship maps you see that represent linkedin or Facebook connections. Gelphi takes that a step further by providing exact calculations.

Data Extraction Tools:
Octoparse:Octoparse is a free and powerful website crawler used for extracting almost all kind of data we need from the website.

Factbook: The World Factbook provides information on the history, people, government, economy, geography, communications, transportation, military, and transnational issues for 267 world entities.

V NEURAL FUZZY SYSTEM

A neuro-fuzzy system is based on a fuzzy system which is trained by a learning algorithm derived from neural network theory. The (heuristically) learning procedure operates on local information, and causes only local modifications in the underlying fuzzy system.

A neuro-fuzzy system can be viewed as a 3-layer feedforward neural network. The first layer represents input variables, the middle (hidden) layer represents fuzzy rules and the third layer represents output variables. Fuzzy sets are encoded as (fuzzy) connection weights. It is not necessary to represent a fuzzy system like this to apply a learning algorithm to it. However, it can be convenient, because it represents the data flow of input processing and learning within the model.

VI.SYSTEM IMPLEMENTATION

Step 1: Neural Fuzzy System Consists of three layers. The input Big Data (Data Set) is given to input layer. In this layer unformatted data is converted into formatted data.

Step 2: The formatted data is given to hidden or middle layer. The set of rules are analysed and learned. The Hadoop tool is used to analyse the data.

Step 3: The Process is executed. The output is displayed. The parameter such as latency, storage usage is calculated. Latency is number of data is executed from large data set.

REFERENCES

[1] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer.MOA: Massive Online Analysis http://moa.cms.waikato.ac.nz/. Journal of Machine Learning Research (JMLR), 2010.

[2] Cascading, http://www.cascading.org/. [11] Facebook Scribe, https://github.com/ facebook/scribe.

[3] U. Kang, D. H. Chau, and C. Faloutsos. PEGASUS:Mining Billion-Scale Graphs in the Cloud. 2012.

[4] J. Langford. Vowpal Wabbit, http://hunch.net/˜vw/,2016.

[5] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson,C. Guestrin, and J. M. Hellerstein. Graphlab: A new parallel framework for machine learning. In Conference on Uncertainty in Artificial Intelligence (UAI), Catalina Island, California, July 2014.

[6] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari. S4:Distributed Stream Computing Platform. In ICDM Workshops, pages 170–177, 2016.

[7] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria,