ISSN 2394-739X

IJRSET

International Journal for Research in Science Engineering and Technology

STRENGTHEN THE INFORMATION CAPITAL BY FASTENING THE DATA QUALITY WITH DEEP LEARNING

[1] Anusuya. K, [2] M. Aramudhan,
[1] Research Scholar, [2] Associate Professor,
[1] Dept. of Computer science, [2] Department of Information Technology,
[1] Mother Teresa Women's University, [2] PKIET,
[1] Kodaikanal, 624 102, India, [2] Puducherry, 609 603, India.

ABSTRACT- In the competitive data driven world, Data & Data Quality (DQ) plays a vital role in each and every step for a successful business as data provider and data consumer. As on today, high quality data becomes Great Asset for Great Decision makers, so most of the companies proactively investing to prevent poor data quality inflows to business, since enterprise systems designed with high data quality ratio in initial days but the data quality level getting tarnished year by year, which challenges the operational efficiency and business functions. In general, data governance team classifies the data fitness for use with data quality algorithms to detect and repair the data errors, which involves high cost and time. To optimize the scenario, we are proposing to integrate the data processing with deep Belief networks (DBN).This proposed approach uses the feature of DBN multi layers to classify the data fitness versus data quality dimensions and auto repair the defects. Classification can be achieved by DBN layer with pre-trained data quality relevant samples and process the big data for all DQ dimensions in single processing irrespective of volume & complex structures.

KEYWORDS- [Data Quality, Deep Learning]

1. INTRODUCTION

Modern days, world becomes informational economy rather than industrial economy and data becomes the intellectual capital for survival and supporting global business functions. Therefore an enterprise needs to capture the happenings as data to extract knowledge and respond to situations faster than competitors for success. Lack of Structure information quality leads to failure in business goals, productivity, cost and customer satisfaction. Research1 shows that 40% of the anticipated value of all business initiatives are never achieved due to poor data quality and also affects operational efficiency, risk mitigation and agility by compromising the decisions made in each of these areas. Data quality issues are injected due to poor data conversion techniques, data from multiple gadgets and operating environments with schema & structural conflicts, system consolidations, manual data entry, changes not captured and loss of expertise. To avoid bad data, the necessity of data quality assurance and management process rises, which is a long-term task requires the continuous efforts across different stages of data transformations with complex data quality procedures like data profiling, cleansing, parsing and duplicate removal. Our approach proposes to use Pre-Trained Deep Belief Networks to process the large datasets for data quality dimensions and checks with quick turnaround time.

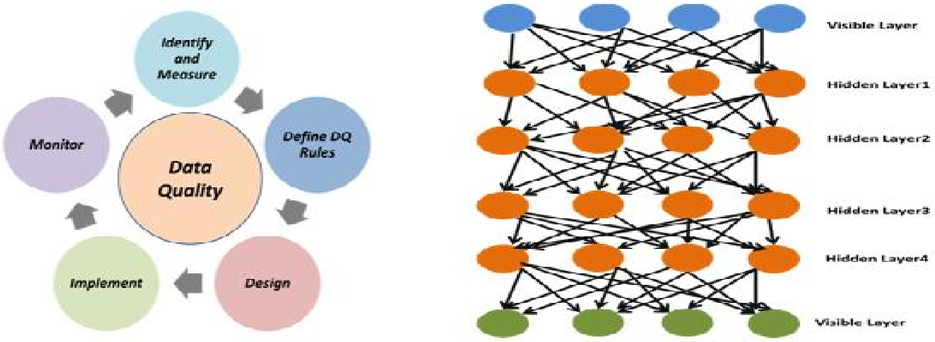
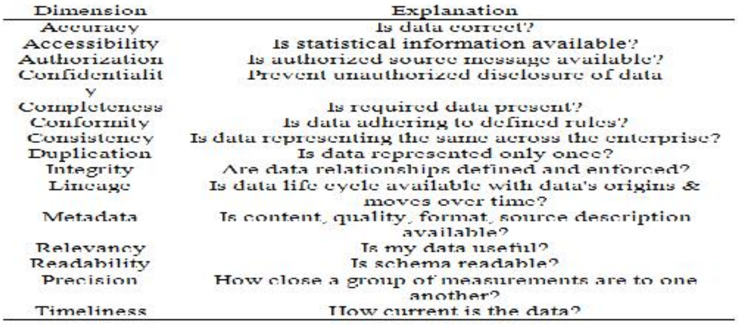


Figure 1- (a)Data Quality Life Cycle (b) DBM Multi-layers

2. DATA QUALITY

In the data driven word, data growing tremendously in Volume, Velocity, Variety and Value as well as data gets collected from multiple data sources with complex data structures (Structured, Semi-structured & Un-Structured). Extracting high quality data[2] from massive volume with complex data structure with in a stipulated time becomes challenge in the dynamic business environment. When poor data quality impacts the business operations, the inevitability of data quality management program ascends and DQ program has data quality process life cycle[3] shown in Fig. 1. (a). This life cycle has five stages to Identify & measure the data quality impedes business objectives, define business-related data quality rules & performance targets, design quality improvement processes that remediate process flaws, Implement quality improvement methods and processes and Monitor data quality against targets. This program shows an overall improvement across the organization and helps to heal and recover the enterprise being dominated by bad data. In general, data quality strategies measure the datum or datasets for "fitness of use". Data quality dimensions [3, 4, 5, 6, 7, 8] classifies the usability of raw data to derive information and data quality, which has been measured with below dimensions.



Dimension	Explanation
Accuracy	Is data correct?
Accessibility	Is statistical information available?
Authorization	Is authorized source message available?
Confidentiality	Prevent unauthorized disclosure of data
Completeness	Is required data present?
Conformity	Is data adhering to defined rules?
Consistency	Is data representing the same across the enterprise?
Duplication	Is data represented only once?
Integrity	Are data relationships defined and enforced?
Lineage	Is data life cycle available with data's origins & moves over time?
Metadata	Is content, quality, format, source description available?
Relevancy	Is my data useful?
Readability	Is schema readable?
Precision	How close a group of measurements are to one another?
Timeliness	How current is the data?

Table 1- Data Quality Dimension.

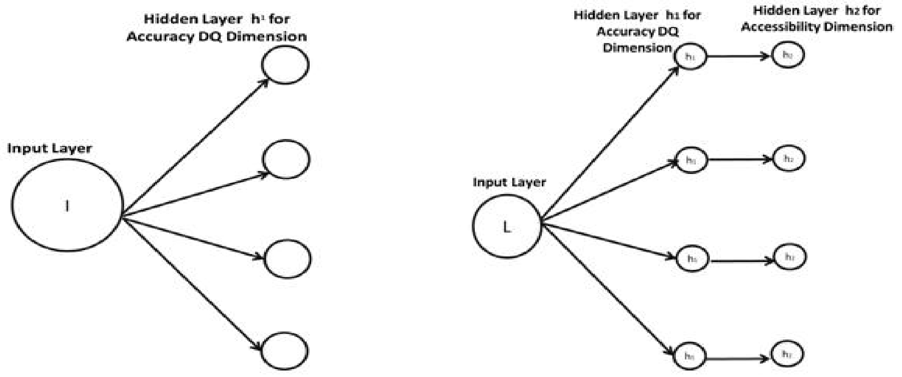


Figure 2- (a) Deep Architecture Level 1 (b) Deep Architecture Level 2

3. DEEP LEARNING

3.1 Restricted Boltzmann machines

Restricted Boltzmann machines (RBMs)[9] are probabilistic graphical models that can be interpreted as stochastic neural networks. The increase in computational power and the development of faster learning algorithms have made them applicable to relevant machine learning problems. RBM has building blocks of multi-layer learning systems called deep belief networks (DBN) for dimensionality reduction, classification, regression, collaborative filtering, feature learning and topic modeling [9].

3.2 Deep Belief Networks (DBN)

Deep Belief Network[10] is generative graphical models consist of multiple layers with hidden units and have connection between layers but not between the hidden nodes in the same layer. In unsupervised way, DBN layers can be trained with set of examples to learn. After this learning step, a DBN can be further trained in a supervised way to perform classification. DBNs can be viewed as a composition of simple, unsupervised networks such as restricted Boltzmann machines . This also leads to a fast, layer-by-layer unsupervised training procedure.

3.3 Pre-train DBN Layer(s)

RBM learn layer by layer by forming stack from bottom up gives rise to a DBN shown in Fig. 1. (b). this stacking procedure improves likelihood of the training data under the composite model and this learning is unsupervised. When applied to classification tasks, the generative pre-training[11] can be followed by or combined with other, typically discriminative, learning procedures that fine-tune all of the weights jointly to improve the performance of the network.

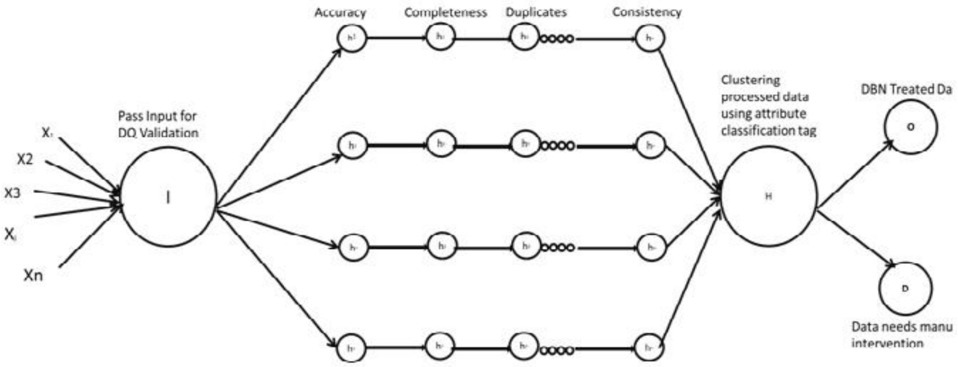


Figure 3- Deep Architecture for Data Quality

4. PROPOSED DATA QUALITY APPROACH

4.1 Problem Statement

In enterprise, the data quality team executes the data quality check on mass volume of data one after the other for each data quality dimensions such as Accuracy, Accessibility, Authorization etc., This data quality check loads the initial data to temporary workspace and profile the data quality vs predefined standardization rule for each DQ dimensions one after the other. This process becomes high cost in terms of time, network and storage, to overcome the limitation this study aims to utilize the benefit of deep belief network’s multiple hidden units which is been pre-trained for n number of data quality dimensions and facilitates the feature of data processing at once.

4.2 Proposed Architecture

In traditional Data Quality processing, Data Quality team pull the operational data from production database to temporary database. While data transmission from the production database to temporary database can also inject issues due to multiple platform, operating system, and database or program development language. Enterprise decision are based on these data, we need to overcome the data quality issues and keep the data quality with higher ratio of reliability, consistency, accuracy and correctness. For data quality processing DBN architecture creation comprises of below steps

- Construct an RBM with an input layer I and a hidden layer h as in Fig.2 (a). Fix sample as input and train the hidden layer h using greedy layer-wise training for data quality Dimension (Eg. Accuracy).
- Stack another hidden layer as in Fig.2 (b). On top of the RBM to form a new RBM and train. Continue to stack layers on top of the network for n data quality dimensions.
- Stack another hidden layer H with single node to process the output.
- Stack visible output layer with nodes O and D for processed output.
- The final Deep Architecture for data quality processing is presented in Fig. 3. Architecture representation is given by $P(I,h^1,h^2 \ldots h^n,H,O,D) = P(I|h^1)P(h^1|h^2) \ldots . . P(h^{n-2}|h^{n-1})P(h^{n-1}|h^n)P(h^n|H)P(H|O)P(H|D)$

5. DATA QUALITY IMPROVEMENT APPROACH USING DEEP BELIEF NETWORK

This study aims to perform data quality analysis for massive volume of data using RBM unsupervised layer-wise pre-training. To summarize the approach, hidden nodes will be pre-trained with relevant samples for each data quality dimensions. This data quality dimensions can be extended to n number of dimensions. Input dataset (x_1, x_2, x_3,…..x_j ,….x_n) passed to visible node I. The input X is the datum with multiple attribute’s shown for Employee entity in Table.2. These attributes assigned with weights w for each data quality dimension and weight w differs for each dimensions based on the data quality fitness level of use.

Name	Weight
Identification Number	0
Name	1
DOB	5
Designation	3
Salary	4
Company	2

Table 2- Employee.

Visible Node I divide the large dataset to smaller subset and passes the input to hidden layers for parallel processing. Each hidden layers (h_1, h_2, $h_3 \ldots h_r$) trained with relevant data samples as per the data quality algorithms (E.g.: Data matching algorithm[12], Data cleaning algorithm[13], Record Linkage algorithms[14]). Error classification Tag used to denote the type of issue detected.

Name	Example
DQ_DIMENSION	Accuracy
ATTRIBUTE_NAME	DOB
ERROR	Invalid date format

Table 3- Classification Error Tag.

Hidden nodes process the dataset and default correction can be made by pre-trained layers. For example date format "3/3/16" can be autocorrected to "03/03/2016". There are scenarios which needs manual intervention will have the classification error tag on. Once processing completed the final hidden node H cluster[15] the dataset into "Treated Data" and "Treatment required data" using classification error tag and sends the output to visible V and D node. Node V has treated data and Node D has data needs manual intervention.

6. APPLICATION AREAS

The proposed approach provides solution for multiple data quality dimensions for Big Data processing using DBN wherever data has higher ratio of data quality issues. Application areas are 1) Financial services sectors to analyze bureau scores, credit card scoring, sentiment analysis and anti-money laundering, regulatory analysis, Stock Market which deals with millions of records with complex data structures. 2) Integrating competitor information using crawler tools, third party data and with own database for market correction in E-Retail Business. 3) Creating regional database across multiple geographies using multilingual data for consumer and pharmaceutical industry. 4) Integration of real time alerts using weather, traffic and global positioning systems (GPS) for Military, Navy, Airforce, public, private and commercial road & air ways systems.

CONCLUSION AND FUTURE WORK

This approach can strengthen the information capital for business enterprise and optimize the data processing time with the help of DBN multiple layers processing unit. Training the DBN layers would be a time consuming process but once layers are pre-trained, the processing unit can classify the data quality dimensions efficiently and faster. Currently this paper, discuss the overall architecture of the Data quality dimensions and identifies the application areas. In future, preparation of pre-training DBN layers for data quality dimensions can be explored and integrated for optimal end-end DBN solution for data quality.

REFERENCES

[1]. Ted Friedman, Michael Smith. Measuring the Business Value of Data Quality. Gartner, 10 October 2011.

[2]. CAI, L and Zhu, Y 2015 The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14: 2, pp. 1-10.

[3]. David Loshin. The Practitioner's Guide to Data Quality Improvement. Elsevier, 2011.

[4]. Carlo Batini, Monica Scannapieca. Data Quality Concepts, Methodologies and Techniques. Springer Publications, 2006.

[5]. Arkady May Danchik. Data Quality Rules. In: Data Quality Assessments. Technical Publications, 2007;57-169.

[6]. Yang W. Lee, Leo L. Pipino, James D. Funk, and Richard Y. Wang. Understanding the Anatomy of Data Quality Problems and Patterns. In: Journey to Data Quality. The MIT Press, 2006; 79-109.

[7]. Shazia shadiq. Computational Aspects of Data Quality. In: Handbook of data quality research and practice. Springer Publications, 2013; 180-320

[8]. Danit McGilvray. Ten steps to Quality Data and Trusted Information. The MIT Press, 2009.

[9]. Asja Fischer, Christian Igel. An introduction to Restricted Boltzmann machines. Volume 7441 of the series Lecture Notes in Computer Science pp 14-36

a, A Beginner's Tutorial for Restricted Boltzmann Machines, http://deeplearning4j.org/restrictedboltzmannmachine.html

[10]. Deep belief network. https://en.wikipedia.org/wiki/Deep_belief_network. Date accessed: 23/03/16.

[11]Li Deng and Dong Yu. Pre-Trained Deep Neural Networks A Hybrid. In: Deep Learning Methods and Applications. Foundations and Trends in Signal Processing. 2014; 240-250.

[12]. Clint Bidlack, Michael P. Wellman. Exceptional Data Quality Using Intelligent Matching and Retrieval. Association for the Advancement of Artificial Intelligence. 2010.

[13]. Kazi Shah Nawaz Ripon, Ashiqur Rahman and G.M. Atiqur Rahaman. A Domain-Independent Data Cleaning Algorithm for Detecting Similar-Duplicates. ACADEMY PUBLISHER, 2010.

[14]. D. Randall Wilson. Beyond Probabilistic Record Linkage: Using Neural Networks and Complex Features to Improve Genealogical Record Linkage. IEEE, 2011

[15]. Mostafa A. SalamaI, Aboul Ella Hassanien, Aly A. Fahmy. Deep Belief Network for Clustering and Classification of a Continuous Data. IEEE, 2011.