ISSN (e): 2250 – 3005 || Volume, 07 || Issue, 11|| November – 2017 ||
International Journal of Computational Engineering Research (IJCER)

An approach to mitigate Veracity issue in Big Data using Regression

C.S.Sindhu[1],Dr. Nagaratna P.Hegde[2]
[1] Research Scholar, JNIAS, Assistant Professor, Global Academy of Technology, Bangalore,
Professor, Vasavi College of Engineering, Hyderabad

ABSTRACT
The rapidly growing field of big data has played a major role in the growth of health care. Big Data analytics facilitates the process of early detection and cure of diseases. The word Big Data brings along with it the various V's associated with it. In this paper, we speak about the veracity issue of Big Data and show how it can lead to better prediction if handled well. We have used age and sex as factors to group patient's based on a standard prescribed by cancer foundation. Modified Moving Average(MMA) is used to predict the missing values. We have used data of a disease called ALS (Amyotrophic Lateral Sclerosis) to prove the correlation between patients. It was found that the results of prediction are better when veracity is handled than when not.
Keywords: Accuracy, Auto Regression Tree, Big Data, Modified Moving Average, Non Linear Regression, Specificity, Veracity

---

Date of Submission: 06-11-2017 Date of acceptance: 17-11-2017

I. INTRODUCTION

Data accumulated every day is growing exponentially [12]Health Sector generates massive amounts of data every second. A lot of effort is going into digitizing this entire data.Data in U.S alone will soon reach zettabyte(1021 gigabytes) and yottabyte(1024 gigabytes) [1]. This large amount of data called Big Data, is defined as data which is variable, of large volume and so complex that it needs advanced techniques to store, analyse and manage this information. [1]Big Data in health care is very difficult to manage not only because of the volume but also because of the speed at which it arrives. [2] Our traditional systems are not capable of handling this large amount of data. Data Storage and Predictive analysis are the two main characteristics of big data. By digitizing these huge volumes, healthcare organizations stand to realize significant benefits [3a]. Some of the benefits are identifying diseases at early stages, detecting fraud in health care at initial stages, and better treatment facilities.

Big data scientists find ample opportunity in this data. Big data is the core of narrative literature. It allows for inference, interpretation, innovation and invention[11].Big data has the potential to discover hidden patterns, improve lives and save costs. This is the reason why it is gaining so much importance. It provides us a platform to extract insights for making better decisions[3-5]. It provides evidence based decision, which organizations need for effective growth. Through this, one could develop better diagnosis and treatments at lower costs. Big data encompasses characteristics such as variety, velocity and veracity [6].The quality of data in the market is growing at 16% above inconsistent data [8, 9]. When dealing with Health Informatics in Big Data, we end up accumulating erroneous or incomplete data(which could be seen with faulty sensors, gene microarrays and dna sequencing) which has to be properly evaluated and dealt with.This paper primarily focuses on the veracity issue of Big Data, and it shows how quality of prediction can be better if veracity is handled effectively.

II. LITERATURE REVIEW

In traditional systems, we get cleaned and manageable data. But many a times, we know that data is uncertain or wrong with many missing values. Veracity goes inline with uncertainty of big data and it is rapidly increasing over the years. Even though there is uncertainty, it still contains useful information.[6a]Uncertainty can be described in terms of latency, inconsistency in data, ambiguities as well as missing values . Uncertainty is classified as expression uncertainty and content uncertainty based on the analysis done on large data. [7]Uncertainty is more complex in textual data since it comes from sources which is highly non reliable and uncertain in content [7] All these challenges question the integrity and authenticity of the raw data and the

results thus obtained. Due to the uncertainty in Big Data, veracity is one of the most significant factors to create value from data.

Veracity deals with the authenticity of data or missing values, i.e the credibility of data. Here we have tried to address the issue of missing values. Studies on data veracity have focused on accomplishing reliability as well as trust of big data. Veracity is the fourth V, according to IBM. It tells us about the uncertainty in the data arriving from various sources. Tatiana Lukoianova[10] have argued that big data has certain properties which affect its quality. Based on this, they have tried to identify what are the quality dimensions for each data type. In [10] they have tried to manage uncertainty by quantifying the levels of content objectivity, truthfulness and credibility. The problems related to veracity become severe when the domain is of health care as any negligence can cost the life of a patient. Hence it becomes very essential to mitigate the issues related to veracity which has resulted in this work.

III. CONTRIBUTION

The data set taken for this experiment is of ALS patients. First we establish a relation matrix by finding the correlation between patients. The demographic variables play a big role in predictive data analysis. It is a known fact that in the medical domain gender has a close correlation with age and can act as a low level classifier. It is because there are certain diseases and clinical conditions which are quite gender-specific.

	P0	P1	P2	P3	P4	P5	P6	P7
Age	[38]	[37]	[36]	[42]	[44]	[51]	[37]	[39]
Sex	[1]	[1]	[1]	[0]	[0]	[1]	[1]	[1]

Figure 1. Data of age and sex of eight patients

In Figure. 1 we have the age of six patients in which the value [1] represents the male gender and [2] represents the female gender. Age and Sex are the two parameters considered in this case to fill in the missing values.

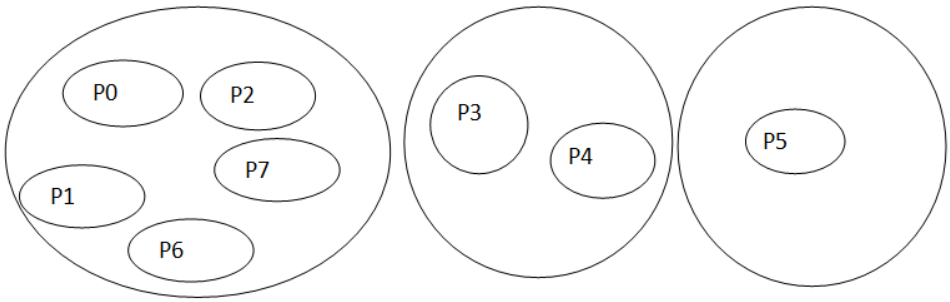

Figure 2. Clusters formed on the basis of age and sex

3.1 Modified Moving Average(MMA)

From the clusters formed in Figure 2, we try to find the correlation between various patients using MMA. In ART, computation is done using Standard Deviation and Mean. In MMA, we have considered only Standard deviation. Moving Average Models can be employed in finding the missing values in large databases. In moving average model, the missing values are calculated based on mean or median values calculated with known data. In the proposed Modified moving average (MMA) model, the missing values are calculated based on standard deviation values.

The standard deviation is computed as

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \qquad \text{----------(1)}$$

μ is the mean of N values .

In Auto regression tree we compute the missing values as

yi=m+$\sum_{i=1}^{n}$ bjyi-1+σ2

where m is the mean , σ is the standard deviation. ART approach works well when the data are linear. Since our data set in nonlinear , this approach does not give good results. Hence we have used the Modified Moving Average, a modification of the moving average method and ART algorithm for nonlinear data. We have considered the parameter of standard deviation from ART to compute our average, hence the proposed method is a variation of ART and Moving Average.

3.2 Results of our approach versus ART and Moving Average

The tables below clearly shows that the modified moving average produces better results when compared to ART and Moving Average methods to fill missing values. Based on this the prediction of the disease also becomes more accurate.

Let the Sample Input Dataset is given in below Table 1.

Table 1: Sample Time series Dataset

SI.No.	*ID*	*Age*	*BP*	*BPD*	*Sugar*	*ALS Total*
1	477518	58	149	85	140	30
2	220672	72	141	87	189	29
3	192626	-	118	78	105	31
4	665671	49	-	85	167	28
5	539837	66	119	-	225	31
6	414036	46	-	38	98	34
7	262418	-	-	-	139	35
8	904513	-	-	-	218	35
9	687046	54	90	-	162	31

If none of the values is given , the filling of records with no known values is not possible with the modified moving average models. Hence, again to fill further, we have introduced the Lagrange polynomial interpolation method.

The missing values are represented by hyphens (-), that are filled through Auto Regressive method is given in below Table 2.

Table 2: Filled missing values through Auto Regressive method

Subject ID	*Age*	*BP*	*BPD*	*Sugar*	*ALSFRS*	*Prognosis*
477518	58	149	85	140	30	1
220672	72	141	87	189	29	1
192626	52	118	109	105	27	0
665671	49	188.33	85	167	28	1
539837	66	119	42.66	225	31	1
414036	46	63.33	38	98	12	0
262418	74	95.46	47.9	139	35	1
904513	43	101.33	67.66	218	35	1
687046	54	90	77	162	31	1

Also, the missing values that are filled through Moving Average method is given in below Table 3.

Table 3: Filled missing values through Moving Average method

Subject ID	*Age*	*BP*	*BPD*	*Sugar*	*ALSFRS*	*Prognosis*
477518	58	149	85	140	30	1
220672	72	141	87	189	29	1
192626	65	118	109	105	27	0
665671	49	130	85	167	28	1
539837	66	119	97	127	31	1
414036	46	125	38	225	12	0
262418	56	122	68	98	35	1
904513	51	124	53	98	35	1
687046	54	90	61	139	31	1

Finally, the missing values that are filled through proposed Modified Moving Average using Lagrange Polynomial interpolation method is given in below Table 4.

Table 4: Filled missing values through Modified Moving Average method

Subject ID	*Age*	*BP*	*BPD*	*Sugar*	*ALSFRS*	*Prognosis*
477518	58	149	85	140	30	1
220672	72	141	87	189	29	1
192626	35	118	109	105	27	0
665671	49	93.33	85	167	28	1
539837	66	75.88	101.82	127	31	1
414036	46	123.98	167.94	225	12	0
262418	46	155.823	168.98	98	35	1
904513	46	84.34	87.954	98	35	1
687046	54	90	134	139	31	1

Moreover, the data filling performance of the proposed and existing methods is evaluated in terms of processing time. The processing time represents the total time required to fill the data. The Processing Time

(seconds) taken during the computation of missing values by varying the number of records from 1000, 2000 to 5000 is given in the below Table 5.

Table 5: Processing Time (min) by varying number of records

Number of Records	Elapsed time (seconds)		
	Auto-regressive Tree (ART)	Moving Average (MA)	Modified Moving Average (MMA)
1000	0.363184	0.240698	0.165237
2000	0.390924	0.264381	0.234192
3000	0.451519	0.32975	0.326254
4000	0.565293	0.442057	0.393132
5000	1.126997	1.077010	0.505075

From above table, it is clear that the processing time is very low for the proposed MMA method.

IV. CONCLUSION

This paper presents a method of filling in the missing values using modified moving average method. This method has taken the concept of calculating missing values using Standard Deviation from ART and produced better results when compared with ART and Moving Average. The reason ART could not produce good results was because it works well for linear data. Since our experiment was on nonlinear data, MMA produced better results. The prediction accuracy was found to be higher in MMA.

ACKNOWLEDGEMENT

We would like to extend our gratitude to the members of PRO-ACT who have developed the data set used in this work. PRO-ACT (Pooled Resource Open-Access ALS Clinical Trials Database) contains about 8500 records of ALS patients whose identity is hidden. In 2011, Prize4Life, in collaboration with the Northeast ALS Consortium, and with funding from the ALS Therapy Alliance, formed the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Consortium. The data available in the PRO-ACT Database has been volunteered by PRO-ACT Consortium members. (https:nctu.partners.org)

REFERENCES

[1]. IHTT: Transforming Health Care through Big Data Strategies for leveraging data in the health care industry; 2013. http://ihealthtran.com/wordpress/2013/03/iht%C2%B2-releases-big-data-research-report-download-today/.
[2]. Frost & Sullivan: Drowning in Big Data? Reducing Information Technology Complexities and Costs for Healthcare Organizations. http://www.emc.com/collateral/analyst-reports/frost-sullivan-reducing-information-technology-complexities-ar.pdf.
[3]. Burghard C:Big Data and Analytics Key to Accountable Care Success.IDCHealth Insights; 2012.
[4]. Ikanow: Data Analytics for Healthcare: Creating Understanding from Big Data. http://info.ikanow.com/Portals/163225/docs/data-analytics-for-healthcare.pdf.
[5]. jStart:"How Big Data Analytics Reduced Medicaid Re-admissions."A jStartCaseStudy; 2012. http://www-01.ibm.com/software/ebusiness/jstart/portfolio/uncMedicaidCaseStudy.pdf.
[6]. Knowledgent:Big Data and Healthcare Payers; 2013. http://knowledgent.com/mediapage/insights/whitepaper/482.
[7]. Menon, S.P. and Hegde, N.P., 2015, January. A survey of tools and applications in big data. In Intelligent Systems and Control (ISCO), 2015 IEEE 9th International Conference on (pp. 1-7). IEEE.
[8]. Claverie-Berge, Isabelle (2012). Solutions Big Data IBM. http://www05.ibm.com/fr/events/netezzaDM_2012/Solutions_Big_Data.pdf
[9]. Schroeck, Michael , Shockley, Rebecca , Smart, Janet , Romero- Morales, Dolores, &Tufano, Peter (2012). Analytics: The real-world use of big data. How innovative enterprises extract value from uncertaindata. www.stthomas.edu/gradsoftware/files/BigData_RealWorldUse.pdf
[10]. G. Beskales, I. F. Ilyas, L. Golab, A. Galiullin, On the relative trust between inconsistent data and inaccurate constraints, ICDE 2013: 541-552
[11]. F. Chiang, R. J. Miller: A unified model for data and constraint repair.ICDE 2011: 446-457
[12]. Veracity Roadmap: Is Big Data Objective, Truthful and Credible? Tatiana Lukoianova and Victoria L. Rubin
[13]. Cronin, Blaise (2013). Editorial. Journal of the American Society for Inform. Science & Technology, 63(3), 435–6.
[14]. Bail, C. A. (2014). The cultural environment: Measuring culture with big data. Theory and Society,43(3–4), 465–482. doi: 10.1007/s11186-014-9216-5

International Journal of Computational Engineering Research (IJCER) is UGC approved Journal with Sl. No. 4627, Journal no. 47631.

C.S.Sindhu An approach to mitigate Veracity issue in Big Data using Regression." IInternational Journal of Computational Engineering Research (IJCER), vol. 7, no. 11, 2017, pp. 51-54.