International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

Modelling Adversarial Risk in Big Data

Boakye Agyemang, MPhil[1,2], Bashiru I. I. Saeed, PhD [3], Albert Luguterah, PhD[2], Samuel Baffoe MSc[1]

[1]Department of Applied Mathematics, Koforidua Technical University, Faculty of Applied Sciences and Technology, Koforidua, Ghana

[2] University for Development Studies, Faculty of Mathematical Sciences, Department of Statistics, Navrongo, Ghana

[3] Tamale Technical University, Faculty of Applied Sciences and Technology, Tamale Ghana

Corresponding Author: Boakye Agyemang; Email: *agyemang.boakye[at]ktu.edu.gh*

Abstract: ***This paper seeks to develop methods to support decisions in relation to adversarial risk analysis of big data by particularly determining some adversarial risk estimators to be derived from big data analysis using the adversarial risk analysis structural equation modelling (ARA-SEM). Data was simulated for one thousand (1000) observations with the results revealing 19 iterative solutions to the latent and measurement models with 16 possessing adversarial risks. The paper recommends the fitting of the ARA-SEM model based on the statistically significant adversarial risk presence as given by the latent and measurement model outcomes.***

Keywords: Risk, Adversarial Risk Measurement Model; Latent Variable; Modelling and Adversarial Risk Analysis (ARA)

1. Background

The assessment and analysis of statistical risk is very crucial and critical in making informed and intelligent decisions particularly in the modern-day scientific era of decision science. This is because decision making has become more complex and scientifically based on empirical evidence emanating from statistical analysis of available data. The essence of paying attention to risk is based on the fact that there has been advancement in information and communication technology coupled with the complexity of the world system through interconnectivity and instrumentation resulting in the type, nature, volume of data changing drastically and evolving into what is termed 'big data'.

However, Reed (2017) posits that big data requires the non-conventional strategies and innovative technologies used by businesses and organizations to capture, manage, process, and make sense of huge amount of data. Ibrahim etal (2015) corroborates that big data analysis requires a combination of analytical techniques and technologies that include new applications to derive benefit or insights from such data.Kantarcioglu and Xi (2016) therefore in discussing the varying approaches developed for adversarial data mining concluded as a two-player game model problem, where the adversary tries to maximize its return and the data miner on the other hand seeking to minimize the misclassification cost. The implications of these results lead to how to choose a set of effective features for adversarial data mining applications. As a result of the techniques, Kantarcioglu and Xi (2016) proposes some attack models against data mining techniques to include a free-range attack model that permits arbitrary data corruption and a restrained attack model that anticipates more realistic attacks that a reasonable adversary would devise under penalties. Banks et al (2015) corroborates the modelling approach of Kantarcioglu and Xi (2016) through the use of adversarial risk analysis (ARA) to model the intentions and strategic behaviour of adversaries in the cyber security domain in particular.

Whilst the above studies seek to model the adversarial intentions of these adversaries by classifying as atwo-player simultaneous game, and in relation to the field of cyber security, its applicability is very limited in the case of big data since the interactions among the opponents in the big data go beyond two player game and equilibrium strategies. This is due to the fact that the huge volumes of data resulting in higher dimensional data or big data requires a new inductive analytical approach since there exist multiple interactions which go beyond two-player game and cannot be modelled as such (Ibrahim et al, 2015). Additionally, these adversaries always have the tenacity to pose threats to systems known as 'intentionality' which is a key factor when analysing all forms risks and threats.

This paper is therefore of the aim to attempt to derive methods to support data analysis in relation to adversarial risk analysis of big data by particularly determining adversarial risk models to be derived from big data analysis using adversarial risk analysis structural equation modelling (ARA-SEM) approach in furtherance to the modelling process, assumptions and proposed model by Boakye et al (2021).

2. Methodology

2.1 Model specification

Based on the assumptions of Boakye et al (2021) there are several events taking place which will lead to several courses of actions, decisions and choices such as in the big data where the events E_1, ..., E_∞ are identifiable with several costs implications:

$$\psi_{Ai}(ai, di) = \int_S g_{Ai}\big(q(ai, di)\big) \times \big(\Sigma_i q_i(ai, di) \int u_{Ai}(ci)\pi_{Ai}(ci|ai, di, E_i)\, dc\big) dq(ai, di) \quad (1)$$

$0 \leq i \leq \infty$,

with the following relations being the modified Nash equilibrium measurement models (MNEMM):

$$\psi_A(a^*, d^*) = max\ \psi_A(a, d^*) \quad (2)$$

$$\psi_D(a^*, d^*) = max\ \psi_D(a^*, d) \quad (3)$$
$$a^*(d) = argmax\ \psi_A(a, d) \quad (4)$$
$$d^* = argmax\ \psi_D(d, a^*(d)) \quad (4)$$
$$d^*_2(d_1, a) = argmax\ \psi_D(d_1, d(d_2)) \quad (5)$$
$$a^*(d_1) = argmax\ \psi_A(d_1, a, d^*_2(d_1, a)) \quad (6)$$
$$d^*_1 = argmax\ \psi_D(d_1, a^*(d_1), d^*_2(d_1, a^*(d_1))) \quad (7)$$
$$d^* = argmax \sum_{a \in A} P_D(c|d) = \int u_D(c)\pi_D(c|a, d)\, d \quad (8)$$

3. Results and Discussions

3.1 ARA-SEM Model Diagram

The ARA-SEM model as proposed in the methodology is first and foremost diagrammatically represented to actually model the relationship between the variables or factors otherwise referred to as intelligent opponents. This graph precisely portrays the relationship that exist in decisions or the course of actions or simply actions taken which results in an associated cost as well as benefits or consequences. The structural equation model aspect models the relationship or interactions based on the structure of the interactions existing between the respective variables, factors or intelligent opponents giving rise to the structural model. In this modelling approach, it is precisely being referred to as multiple or multivariate structural influence model diagram (MSIMD). This is so due the fact that there are several or multiples of these intelligent opponents in the big data. It is a modification to that of Rios (2009) for adoption and use in most decision analysis and artificial intelligence, neural networks or big data environment in adversarial risk analysis. The proposed diagram is therefore given in the Figure 1.

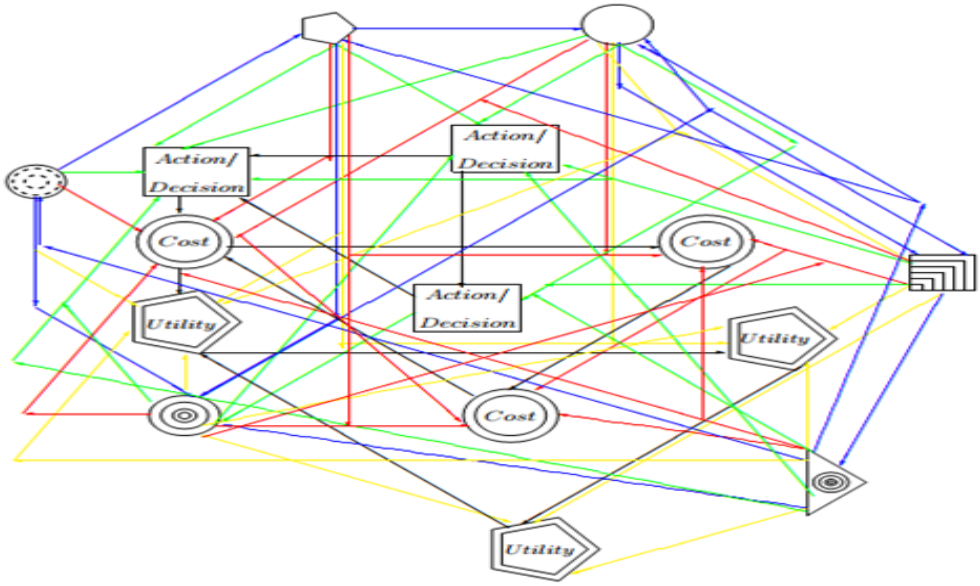

Figure 1: Multiple or multivariate structural influence model diagram (MSIMD)

The symbols denote various components in the big data system where several variables and factors make up the data ecosystem prone to risk referred to as adversarial risk which result from the interactions that takes place among the various agents, factors or variables. All symbols which are unlabelled in the figure represent the various agents or factors which are connected to one another with blue lines, indication the interactions existing between the respective agents. The interactions that exist among the agents come through the form of games which basically yields three unique components. Every agent is an opponent to one another referred to as intelligent opponents with the characteristic of hiding their interaction, action, decision or influence with or to one another from one another. These components include decision or action that each agent makes which comes with a cost and a benefit or value represented by the nodes rectangle, double circle and as well as the diamond shapes. These three components are the attributes associated with each agent, or factor or variable or opponent and are played based on a strategy. They are uniquely connected with separate coloured lines for easy identification purposes. Specifically, all actions or decisions are connected with dark or black lines, the cost lines depicted with red lines and the benefit or values nodes connected with yellow lines. The multiple or multivariate structural influence model diagram (MSIMD) is more complicated in structure and components as compared to the influence diagram developed by Insuah (2009). This therefore makes it more applicable to varying sets of higher dimensional data.

3.2The Adversarial Risk Identification

Subsequent to the determination of the multiple structural influences model diagram are the parameter, variable or opponent or agent interrelationships. The opponent or factor correlations are important in the determination and estimation of the presence of adversaries with a given interaction between and among intelligent opponents. These interactions are examined using the correlations as contained in the Table 2 below. From the results presented in Table 2, only two factors on the other hand relates perfectly among themselves, with a coefficient of correlation of 1. These included wind and arrow, however, a few other variables indicated a moderately strong relationship with majority indicating weak correlations thereby implying a weak or no

relationship. The assumption here is that the adversaries usually hide their intentions and so are depicted by lower correlations (Whirtman et al, 2015). This is clearly a sign of the presence of adversaries since they either hide their actions or interactions from their opponents. The interactional relationship is further examined descriptively in a purely quantitative approach using the results projected subsequently in Table 2.On the contrary, higher correlations among opponents is not risky or does not indicate the presence of adversarial risk particularly in the absence of other statistics or parameters.

Out of a total of 171 interactional coefficient estimates, it can be seen that the minimum value is r = -0.001 whilst the maximum value is r = 1.00 representing a very weak negative and perfect positive relationships respectively. Again, ninety (90) estimators shows a very weak negative interactional relationships with interactional relationship coefficients ranging between r = -0.001 to -0.082. Seventy-six (76) estimators shows a very weak positive relationship also with interactional relationship coefficients ranging from r = 0.002 to r =0.082.Furthermore, three (3) estimates (r = 0.235, r = 0.274 and r = 0.276) shows weak positive relationship, with two (2) other estimators with values (r = 0.704 and r = 0.968) indicating very strong positive relationship and only one (1) estimate (r = 1.00) indicating a perfect positive relationship. This implies clear situations of both hidden and exposed interactional relationships among the opponents.

These results of interactional effects shows that only five (5) estimates which include (r = 0.704; r = 0.968 and r = 1.00) out of a total of 171 interactional effect estimates are consistent with the estimated effects sizes of Hair et al (2010) who proposes that as the sample size increases, the effect as measured by correlation will correspondingly increase and vice versa. Specifically, they found that a larger sample size of 200 at significant level of 0.05 will lead to effect sizes of small of 0.516 and moderate of 0.998 respectively and at a significance level of 0.01 will lead to an effect size of small of 0.284 and moderate of 0.992 respectively. On the basis of the result as thoroughly discussed here, the exact risk functions are then obtained as indicated below. The implication of the interactional effect estimates despite the larger size of the samples used underscore the presence of adversarial relationship since the adversaries are intelligent opponents and for that matter tend to hide their actions or interactions from each other and one another.

Table 2: Intelligent opponents' empirical correlations (between the variables)

	arrow	*under*	*interior*	*theta*	*Amb*	*slice*	*delta*	*Pi*	*height*	*Nu*	*night*	*dataset*	*length*	*volume*	*sales*	*wind*	*adverts*	*rho*	*alpha*
arrow	1.000																		
under	0.009	1.000																	
interior	-0.032	-0.044	1.000																
theta	-0.006	-0.057	0.019	1.000															
Amb	-0.002	-0.004	0.019	-0.021	1.000														
Slice	-0.082	-0.017	-0.022	-0.030	0.070	1.000													
delta	-0.001	-0.025	-0.015	0.042	0.021	0.009	1.000												
Pi	0.007	-0.034	-0.006	0.044	-0.049	-0.007	0.012	1.000											
height	-0.076	-0.025	-0.020	-0.029	0.070	0.968	0.008	-0.008	1.000										
Nu	0.002	-0.007	0.033	0.704	-0.004	-0.012	0.035	0.044	-0.016	1.000									
night	-0.017	-0.032	0.039	-0.023	-0.043	-0.013	-0.047	0.018	-0.014	-0.026	1.000								
dataset	-0.008	-0.009	0.235	-0.013	-0.043	0.024	-0.032	0.007	0.026	0.012	-0.015	1.000							
length	0.009	0.021	-0.001	0.082	-0.004	0.032	0.026	-0.016	0.039	0.015	-0.006	0.004	1.000						
volume	0.010	0.080	0.014	0.019	-0.003	0.009	0.007	-0.026	0.010	-0.035	-0.016	0.014	0.001	1.000					
Sales	0.002	-0.084	-0.019	-0.013	-0.006	0.019	-0.051	0.082	0.019	0.003	0.276	-0.003	-0.044	-0.020	1.000				
Wind	1.000	0.009	-0.032	-0.006	-0.002	-0.082	-0.001	0.007	-0.076	0.002	-0.016	-0.008	0.009	0.010	0.003	1.000			
Adverts	-0.013	-0.021	0.011	0.028	-0.020	0.025	0.120	-0.026	0.016	0.061	0.039	-0.046	-0.008	0.051	0.026	-0.013	1.000		
Rho	0.028	-0.030	0.002	0.063	0.054	0.041	0.044	0.274	0.040	0.062	-0.024	-0.033	-0.011	-0.014	-0.033	0.028	0.004	1.000	
alpha	0.012	-0.023	-0.020	-0.028	-0.016	-0.018	-0.037	-0.044	-0.021	-0.032	0.015	-0.005	0.017	-0.047	-0.018	0.012	-0.050	0.006	1.000

Table 3A: Simulation results for the optimized interactional relationships for adversarial risk models

Relation/Function	$\bar{\Psi}$	$\bar{\Psi}^2$	$P\bar{\Psi}^2$
M1, $K\ [X_1 = f(\bar{X}_1, X_2 \ldots\ldots X_{19})]$	0.120	0.140	0.704
M2, $K\ [X_2 = f(\bar{X}_2, X_1 \ldots\ldots X_{19})]$	1.000	1.000	0.000
M3, $K[X_3 = f(\bar{X}_3, X_1 \ldots\ldots X_{19})]$	0.165	0.027	0.074
M4, $K\ [X_4 = f(\bar{X}_4, X_1 \ldots\ldots X_{19})]$	0.259	0.067	0.001
M5, $K\ [X_5 = f(\bar{X}_5, X_1 \ldots\ldots X_{19})]$	0.714	0.509	0.000
M6, $K[X_6 = f(\bar{X}_6, X_1 \ldots\ldots X_{19})]$	0.136	0.018	0.031
M7, K $[X_7 = f(\bar{X}_7, X_1 \ldots\ldots X_{19})]$	0.969	0.938	0.000
M8, K $[X_8 = f(\bar{X}_8, X_1 \ldots\ldots X_{19})]$	0.172	0.103	0.041

M9, K [$X_9 = f(\bar{X}_9, X_1 \ldots\ldots X_{19})$]	0.304	0.093	0.000
M10, K [$X_{10} = f(\bar{X}_{10}, X_1 \ldots\ldots X_{19})$]	0.969	0.938	0.000

Table 3B: Simulation results for the optimized interactional relationships for adversarial risk models

Relation/Function	$\bar{\Psi}$	$\bar{\Psi}^2$	$p\bar{\Psi}^2$
M10, K [$X_{10} = f(\bar{X}_{10}, X_1 \ldots\ldots X_{19})$]	0.969	0.938	0.000
M11,K [$X_{11} = f(\bar{X}_{11}, X_1 \ldots\ldots X_{19})$]	0.711	0.506	0.000
M12, K [$X_{12} = f(\bar{X}_{12}, X_1 \ldots\ldots X_{19})$]	0.292	0.085	0.000
M13, K [$X_{19} = f(\bar{X}_{19}, X_1 \ldots\ldots X_{18})$]	0.225	0.065	0.000
M14, K [$X_{14} = f(\bar{X}_{14}, X_1 \ldots\ldots X_{19})$]	0.131	0.017	0.522
M15, K [$X_{15} = f(\bar{X}_{15}, X_1 \ldots\ldots X_{19})$]	0.142	0.020	0.328
M16, K [$X_{16} = f(\bar{X}_{16}, X_1 \ldots\ldots X_{19})$]	0.312	0.097	0.000
M17, K [$X_{17} = f(\bar{X}_{17}, X_1 \ldots\ldots X_{19})$]	1.000	1.000	0.000
M18, K [$X_{18} = f(\bar{X}_{18}, X_1 \ldots\ldots X_{19})$]	0.179	0.032	0.020
M19, K [$X_{19} = f(\bar{X}_{19}, X_1 \ldots\ldots X_{18})$]	0.303	0.092	0.000

Table 3C: Simulation results for the adversarial risk estimates

Risk model	ROC	RS	CPS	Result/Implication/Decision
M1, K[$X_1 = f(\bar{X}_1, X_2 \ldots X_{19})$]	0.14	289.345	>12	Not Supported, intolerable, adverse risk
M2, K[$X_5 = f(\bar{X}_5, X_1 \ldots X_{19})$]	1.00	6.409	<12	Supported, tolerable, no adverse risk
M3, K[$X_3 = f(\bar{X}_3, X_1 \ldots X_{19})$]	0.027	1.017	1-5	Not Supported, negligible, no adverse risk
M4, K[$X_4 = f(\bar{X}_4, X_1 \ldots X_{19})$]	0.067	5.1E+10	>12	Supported, tolerable, adverse risk
M5, K[$X_2 = f(\bar{X}_2, X_1 \ldots X_{19})$	0.509	9119.11	>12	Supported, intolerable, adverse risk
M6, K[$X_6 = f(\bar{X}_6, X_1 \ldots X_{19})$]	0.018	142.704	>12	Supported, high priority, adverse risk
M7, K[$X_7 = f(\bar{X}_7, X_1 \ldots X_{19})$]	0.938	991.723	>12	Supported, intolerable, adverse risk
M8, K[$X_8 = f(\bar{X}_8, X_1 \ldots X_{19})$]	0.103	54.183	>12	Supported, critical, adverse risk
M9, K[$X_9 = f(\bar{X}_9, X_1 \ldots X_{19})$]	0.093	4072.2	>12	Supported, critical, adverse risk
M10, K[$X_{10} = f(\bar{X}_{10}, X_1 \ldots X_{19})$]	0.938	1811033	>12	Supported, critical, adverse risk
M11, K[$X_{11} = f(\bar{X}_{11}, X_1 \ldots X_{19})$]	0.506	1.2E+07	>12	Supported, intolerable, adverse risk
M12, K[$X_{12} = f(\bar{X}_{12}, X_1 \ldots X_{19})$]	0.085	1.862	1-5	Supported, acceptable, no adverse risk
M13, K[$X_{13} = f(\bar{X}_{13}, X_1 \ldots X_{19})$]	0.065	3710.8	>12	Supported, critical, adverse risk
M14, K[$X_{14} = f(\bar{X}_{14}, X_1 \ldots X_{19})$]	0.017	0.979	<12	Not Supported, tolerable, no adverse risk
M15, K[$X_{15} = f(\bar{X}_{15}, X_1 \ldots X_{19})$]	0.02	439.413	>12	Not Supported, critical, adverse risk
M16, K[$X_{16} = f(\bar{X}_{16}, X_1 \ldots X_{19})$]	0.097	1.706	1-5	Supported, acceptable, no adverse risk
M17, K[$X_{17} = f(\bar{X}_{17}, X_1 \ldots X_{19})$]	1.00	10.312	<12	Supported, tolerable, no adverse risk
M18, K[$X_{18} = f(\bar{X}_{18}, X_1 \ldots X_{19})$]	0.032	33.089	>12	Supported, critical, adverse risk
M19, K[$X_{19} = f(\bar{X}_{19}, X_1 \ldots X_{18})$]	0.092	5.4E+22	>12	Supported, intolerable, adverse risk

The Table 3A, 3B and 3C display the Modified Nash equilibrium measurement models as the main criteria for model identification. It indicates that a total of thirteen (13) models had adversarial risk present out of a total of nineteen (19) measurement models assessed in Table 3C. The models with adversarial risk have been coloured red as indicated in

the last column (Result/Implication/Decision) of the Table 3C upon the assessment of the risk associated with the various interactions. The remaining six (6) functions coloured in blue have risk associated but not adversaries. This implies that the thirteen (13) models will then have to be used in the modelling of their structure. This result is supported by Wirthmann et al (2015) who concluded that priority for actions should be put on the critical risks or those which are likely to happen and have major extreme impact on the objectives of the organisation.

4. Findings and Conclusion

The following are the main findings derived from the study. Firstly, the results indicate that a total of thirteen (13) models had adversarial risk present out of a total of nineteen (19) measurement models that assessed. The models with adversarial risk had smaller or weaker interactional effects as measured by their correlations, but weaker values of returns or relative outer gain (ROC), higher values of relative risk comparable to Sharpe Ration (SR) and as well as corresponding higher levels of risk (CPS). The paper concludes based on the findings derived from the results that the adversarial risk models identified serves as the basis for the development of adversarial as decision regarding identified risk to critically to inform intelligent decisions regarding the use of big data.

The paper recommends the use of only the latent and measurement models with significant adversarial risk presence in the modelling of process of the ARA-SEM model, however, their fitness levels must be ascertained particularly in situations where the dimensionality of the data may be unreasonably higher and may not necessitate the use of all risk models identified.

References

[1] Arce, D., and Sandler, T. (2007), ‘‘Terrorist Signalling and the Value of Intelligence,’’ British Journal of Political Science, 37, 573–586. Arnold Publishing.
[2] Banks, D., and Anderson, S. (2006), ‘‘Game Theory and Risk Analysis in the Context.
[3] Dedić, N.; Stanier, C. (2017). "Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery". 285. Berlin; Heidelberg: Springer International Publishing.
[4] Ibrahim; Targio Hashem, Abaker; Yaqoob, Ibrar; Badrul Anuar, Nor; Mokhtar, Salimah; Gani, Abdullah; Ullah Khan, Samee (2015). "big data" on cloud computing: Review and open research issues". Information Systems. 47: 98–115.
[5] Kakushadze, Z.andYu, W. (2016). Statistical Risk Models. Quantigic Solutions LLC. Centre for Computational Biology, Duke- NUS Medical School 8 College Road, Singapore.
[6] Kantarcioglu, M., Xi, B., and Clifton, C. (2016). Adversarial Data Mining: Big Data Meets Cyber Security. Vienna, Austria ACM ISBN 978-1-4503-4139-4/16/10.
[7] Medvedev, V., Kurasova, O., Bernatavi, J., Treigys, P., Marcinkevi, V., Dzemyda, G. (2017). A new web-based solution for modelling data mining processes. Simulation Modelling Practice and Theory 76 (2017) 34–46. Elsevier.
[8] Rasoolimanesh, S. M., Ringle, C. M., Sarstedt, M., Olya, H. (2021). The combined use of symmetric and asymmetric approaches: partial least squares-structural equation modeling and fuzzy-set qualitative comparative analysis. International Journal of Contemporary Hospitality Management. Emerald Publishing Limited, Pp. 0959-6119.
[9] Reed, J. (2017), Data Analytics: Applicable Data Analysis to Advance Any Business Using the Power of Data Driven Analytics.
[10] Rios, D. I., Rios, J. & Banks, D. (2009) Adversarial Risk Analysis, Journal of the American Statistical Association, 104:486, 841-854, DOI: 10.1198/jasa.2009.0155.
[11] White, T. (2015). Hadoop: The definitive guide (4th, revised & updated ed.). Sebastopol, CA: O'Reilly Media.
[12] Wirthmann A, Karlberg, M., Kovachev B., Reis F. (2015). Structuring risks and solutions in the use of big data sources for producing official statistics –Analysis based on a risk and quality framework. Working Paper, CONFERENCE OF EUROPEAN STATISTICIANS Workshop on Statistical Data Collection: Riding the Data Deluge 29 April – 1 May 2015, Washington D.C., United States of America.