Neuro-fuzzy risk prediction model for computational grids

Neuro-Fuzzy Risk Prediction Model for Computational Grids

Sara Abdelwahab[1,2], Varun Kumar Ojha[3], Ajith Abraham[3,4]

[1]Faculty of Computer Science & Information Technology, Sudan University of Science and Technology, Khartoum, Sudan
[2] Computer Science & Information college, princess Norah BintAbddulrahmanUniversity, Riyadh, Saudi Arabia
saraabdelghani@gmail.com,saabdelghani@pnu.edu.sa
[3]IT4 Innovations - Center of excellence, VSB -Technical University of Ostrava, Czech Republic
[4]Machine Intelligence Research Labs (MIR Labs), Washington, USA
varun.kumar.ojha@vsb.cz,ajith.abraham@ieee.org

Abstract. Prediction of risk assessment is demanding because it is one of the most important contributory factors towards grid computing. Hence, researchers were motivated for developing and deploying grids on diverse computers, which is responsible for spreading resources across administrative domains so that resource sharing becomes effective. Risk assessment in grid computing can analyze possible risks, that is, the risk of growing computational requirements of an organization. Thus, risk assessment helps in determining these risks. In this, we present an adaptive neuro-fuzzy inference system that can predict the risk environment. The main goal of this paper is to obtain empirical results with an illustration of high performance and accurate results. We used data mining tools to determine the contributing attributes to obtain the risk prediction accurately.

Keywords: risk assessment, prediction, adaptive neuro-fuzzy system

1 Introduction

Many risk factors are associated with grid computing that threatens security measures [1]. In this work, we applied a hybrid approach to model risk in the computational grid. We used an adaptive neuro-fuzzy inference system (ANFIS) for modelling risk prediction in a computational grid environment. ANFIS is a fuzzy inference system learned using neural network type learning methods. Using a hybrid learning procedure, ANFIS can construct an input-output mapping based on both human-knowledge as fuzzy '*if-then* rules and approximate membership functions from the stipulated input-output data pairs. ANFIS learning employs a hybrid method consisting of back-propagation for tuning the parameters associated with input membership and least-squares-estimation for tuning the parameters associated with the output membership functions [2].

Researchers excessively used ANFIS in many significant research problems, such as industry, financial, weather-prediction, health, etc. Beghdad et al.[3] used combination of ANFIS and clustering process applied on the CPU Load time series to predict values of CPU load. Their proposed model achieves significant improvement and outperforms the existing CPU load prediction models reported in the literature. In [4], ANFIS was used to predict average air temperature, while authors in [2] used ANFIS to predict roughness surface in ball-end milling aluminium.

The primary contributions of our research work are to conduct a pre-study based on ANFIS that can provide an insight of predicting the risk environment. We organize this paper as follows: Section 2 provides reviews on previous and related work; Section 3 presents the proposed methodology; the illustration of experimental results and discussion of the work presents in Section 4; Section 5 provides discussion followed by conclusions in Section 6.

2 Related Research

Many hybrid approaches have been applied to predict the risk assessment of grid computing and achieved acceptable results. Risk assessment in grid computing was presented at two layers - resource provider (RP) and Broker by a project of AssessGrid (AGP) [5]. In the beginning, the risk modelling of the AGP project was conceded at RP considering the probabilistic as well as the possibility approaches. The risk assessment at the RP level in AGP was accomplished by the Bayesian model and provided the values of risk at the node level. This approach followed the same context as that of the node as the work proposed in [6]. Zadeh [7] proposed the possibility modelling, which Broker level was intended to present a Broker. The Broker was introduced to facilitate the end-user to communicate as well as negotiate with RP. Also, the level was designed so that it can make a selection of the relevant RP among others [8, 9].

In this paper, we do not focus on Broker level work in the AGP. The risk modelling is accountable in AGP at the node level rather than the component level. In addition, the risk models of AGP do not reflect any insight into the grid failure data. In [10], the authors used various reliability models to assess and evaluate on the basis of assuming Weibull distribution as the best-fit model. However, their work suffered a limitation of the requirement of aggregation at both the levels of component and node level. Further, this concept was enhanced to improve reliability within the grids using a stochastic model that extracted grid-trace logs and thus enhanced the job resubmission strategy. Moreover, these works do not address the component-level risk assessment in grids, where the components could be either the disks, CPU, computer software, computer memory, etc. The types of grids are also not classified based on risk assessment, whether the grids are replaceable or repairable. In [11], authors address the problem of risk assessment in computational grid considering security aspect, while most of the earlier proposed model addressed risk assessment in grid considering resource failure aspect. In this work, we extend the work provided in [11] by using a hybrid ANFIS method to predict risk in the computational grid environment.

3 Proposed Methodology

We divided our work into five phases. The description of these phases as follows:

3.1 Phase One – ANFIS Structure

The fuzzy inference system uses fuzzy logic for formulating a nonlinear mapping from input to output, where this system has three parts. (a) A rule-base containing fuzzy rules are selected. (b) Database, which defines membership functions applied for the fuzzy rules. (c) A logical system performing the way of inference based on the rules and facts [2].

The ANFIS network contains five layers, where each layer contains several nodes described by the node function. In the first layer, every node is an adaptive node with a node function such as a trapezoidal membership function or a Gaussian membership function. In the second layer, each node multiplies incoming signals, and the output is the product of all the incoming signals, where each node output represents the firing strength of a rule. In the third layer, each node calculates the ratio of the rules firing strength to the sum of all rules firing strengths. The normalized firing strengths are the output from this layer. In the fourth layer, each node calculates the contribution of the rule to the overall output. In the fifth (final) layer, the single node calculates the final output as the summation of all input signals [4].

3.2 Phase Two –Dataset Selection

In this phase, the dataset is obtained by simulating the grid-computing environment and select the risk factors that threaten the grid-computing environment [11]. The dataset consists of 20 risk factors, and 1951 instances are used to predict the risk output for the grid computing environment (Table 1).

3.3 Phase Three – Selection of the best input model variables

Feature selection is a preprocessing step that reduces dimensionality from a dataset to improve prediction performance. Feature selection can be viewed as a search problem, where searching of a subset from the search space in which each state represents a subset of the possible features. To avoid the high computational cost and enhance the prediction accuracy, irrelevant input features are reduced from the dataset before constructing the prediction model. We used a correlation feature selection (CFS) subset evaluator feature selection algorithm to search for the best-input model variables [12]. In this research, CFS was used along with the ANFIS to evaluate the merit of feature subsets.

Table 1. Grid Risk Assessment Factors

Risk Factor	Definition	Ref.
Services Level Agreement Violation (SLAV)	SLA represents an agreement between a service user and a provider in the context of a particular service provider.	[13]
Cross Domain Attack (CDA)	CDA in which the attacker compromises one site and can then spread his attack easily to the other federated sites.	[14]
Job Starvation (JS)	In JS, stranger job scheduled on the host uses local (host) resources.	[15]
Resource Failure (RF)	It is a failure if: (i) resource stops because of resource crash; (ii) available resources do not meet the minimum levels of QoS.	[16]
Resource Attacks (RA)	It is illegal to use host resources by an attacker.	[17]
Privilege Attack (PA)	User may gain the excess privilege of accessing the command shell. If grid computing allows access to command shell using predefined scripts.	[17]
Confidentiality Breaches (CB)	Unauthorized, unanticipated, or unintentional disclosure could result in loss of public confidence or legal action against the organization.	[18]
Integrity Violation (IV)	Integrity refers to the trustworthiness of data or resources, and it is usually phrased in terms of preventing improper or unauthorized change	[18]
Distributed Denial of Services (DDoS)	DoS attacks involve sending a large number of packets to a destination to prevent legitimate users from accessing information or services.	[19]
Data Attack (DA)	In grid security, DA is a scheme in which malicious code is embedded in innocuous-looking data, which (when executed by a program) plays out the intended destructive results.	[17]
Data Exposure (DE)	DE is another side of widespread connectivity in which (while improving productivity) makes it easier to obtain unauthorized to sensitive data	[17]
Credential Violation (CV)	Credentials are tickets or tokens used to identify, authorize, or authenticate a user. Comprise CV causes theft of user credentials.	[18]
Man in the Middle Attack (MMA)	MMA is an attack where the attacker secretly relays and possibly alters the communication between two parties.	[15]
Privacy Violation (PV)	PV is the interference of a person's right to privacy by various means, such as showing photos in public.	[20]
Sybil Attack (SA)	In Sybil attacks, few entities fake multiple identities. So it is a concern for the systems that rely upon implicit certification.	[15]
Hosting Illegal Content (HIC)	This can be done by exploiting the leased nodes.	[17]
Stealing Input or Output (SIO)	It is a way to steal the data received by the system or to steal data sent from it.	[17]
Shared Use Threats (ShUTh)	Incompatibility between the attributes of grid users and conventional users causes ShUTh. Hence, no strict separation between participants.	[17]
Stealing or altering the Software (SS)	SS is caused by unauthorized means of entering or altering data, false data, unauthorized data, or unauthorized instruction to a system.	[17]
Policy Mapping (PM)	Multiple administrative domains with multiple policies cause difficulty to users to map different policies across the grid	[18]

3.4 Phase Four – Investigation of the effectiveness of data splitting

The fourth phase is partitioning data to investigate the effectiveness of data splitting. We randomly split our dataset into training and testing set as follows:

A: Split 60 % training, 40% testing
B: Split 70 % training, 30% testing
C: Split 80 % training, 20% testing
D: Split 90 % training, 10% testing

3.5 Phase Five - development and configuration of ANFIS

In this phase, the ANFIS model was constructed using grid partitioning, and the membership function and consequent parameters were tuned using a hybrid learning process for 100 epochs. We used different membership functions to represent each input variable [21]. In this work, we used:
Trapezoidal membership function (Trapmf): Trapezoidal curve is a function of a vector x and depends on four parameters *a, b, c,* and *d*, as given by:

$$f(x,a,b,c,d) = \max\left(\min\left(\frac{\mathrm{x}-\mathrm{a}}{\mathrm{b}-\mathrm{a}}, 1, \frac{\mathrm{d}-\mathrm{x}}{\mathrm{d}-\mathrm{c}}\right), \mathrm{o}\right)$$

Parameter *a* and parameter *d* locate the "feet" of the trapezoid, and parameters *b* and *c* locate the "shoulder."

Triangular membership function (Trimf): The triangular curve is a function of a vector *x* and depends on three scalar parameters, *a, b*, and *c*, given by:

$$f(x,a,b,c) = \max\left(\min\left(\frac{\mathrm{x}-\mathrm{a}}{\mathrm{b}-\mathrm{a}}, \frac{\mathrm{c}-\mathrm{x}}{\mathrm{c}-\mathrm{b}}\right), \mathrm{o}\right)$$

Parameter *a* and parameter *c* locate the "feet" of the triangle, and parameter *b* locates the peak.

Generalized bell function (Gbell): Depends on three scalar parameters, *a, b,* and *c*, given by:

$$f(x,a,b,c) = \frac{1}{1+\left|\frac{x-c}{a}\right|^{2b}}$$

Where parameter *b* is usually positive, and parameter *c* locates the centre of the curve.
Gaussian membership function (Gaussmf): The symmetric Gaussian function depends on two parameters *σ* and *c* as given by:

$$f(x,\sigma,c) = e^{-\frac{(x-c)}{2\sigma^2}}$$

4 Experimental Dataset and Empirical Design

Using different feature selection technique, eight different sub-datasets were obtained [11]. These datasets were generated from our original dataset that has 20 risk factors attributes. Table 2 illustrates the number of attributes in each dataset and summarizes the search method.

Table 2. Attributes Selection Methods

Dataset	Evaluator	Search method	Selected Attributes	Total
Original dataset			SLAV, CDA, JS, RF, RA, PA, CB, IV, DDoS, DA, DE, CV, MMA, PV, SA, HIC, SIO, ShUTh, SS, PM	20
1	RelifF Attribute Evaluation	Ranker	DDoS, PM, DE, SA, ShUTh, HIC, CV, RA, SIO, CDA, RF, SLAV, JS, MMA, SS, PA, PV, IV, CB, DA	20
2	Reliff Attribute Evaluation	Ranker	DDoS, PM, DE, SA, ShUTh, HIC, CV, RA, SIO, CDA, RF, SLAV, JS, MMA, SS, PA, PV, IV	18
3	Reliff Attribute Evaluation	Ranker	DDoS, PM, DE, SA, ShUTh, HIC, CV, RA, SIO, CDA, RF, SLAV, JS, MMA, SS	15
4	Reliff Attribute Evaluation	Ranker	DDoS, PM, DE, SA, ShUTh, HIC, CV, RA, SIO, CDA, RF, SLAV	12
5	Reliff Attribute Evaluation	Ranker	DDoS, PM, DE, SA, ShUTh, HIC, CV, RA, SIO	9
6	CFS Subset Eval	Evolutionary Search	SLAV, JS, RA, CV, HIC, SIO	6
7	CFS Subset Eval	Best first search backward	CV, HIC, SIO	3
8	CFS Subset Eval	Exhaustive search	RA, CV, HIC	3

In this study, we extended the work reported in [11], which includes three attributes. The empirical result shows that the prediction algorithm required the least number of attributes (3 attributes only out of 20 attributes) to achieve high performance. In this work, to verify the efficiency of the proposed method, we used three features: CV, HIC, and SIO.

5 Result and Discussion

Different ANFIS parameters were tested as training parameters to maximize the prediction accuracy to achieve the experimental result. Figure 1 illustrates the membership function (TriMF) as input functions.

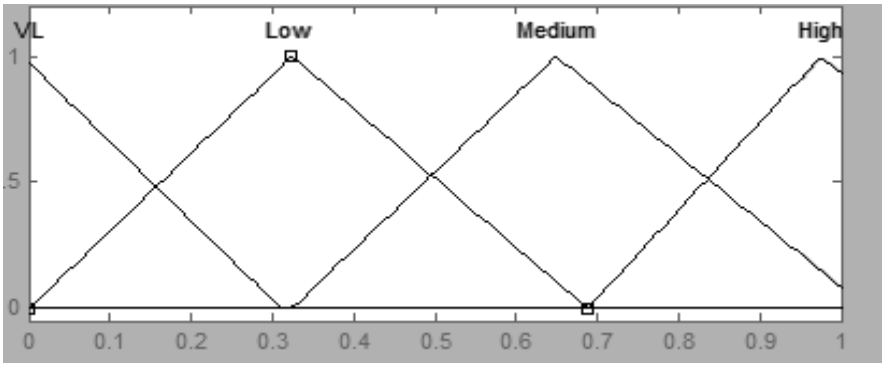

Fig. 1. Triangular-shaped membership function for the input variables

Total 64 fuzzy 'if-then rules' were used to build the fuzzy inference system, and some of the if-then rules are illustrated below.

1. If (CV is VL) and (HIC is VL) and (SIO is VL), then (RO is VL) (1)
2. If (CV is VL) and (HIC is VL) and (SIO is Low), then (RO is VL) (1)
3. If (CV is VL) and (HIC is VL) and (SIO is M), then (RO is VL) (1)
4. If (CV is VL) and (HIC is VL) and (SIO is High), then (RO is VL) (1)
17. If (CV is Low) and (HIC is VL) and (SIO is VL), then (RO is Low) (1)
18. If (CV is Low) and (HIC is VL) and (SIO is Low), then (RO is Low) (1)
37. If (CV is Medium) and (HIC is Low) and (SIO is VL), then (RO is Medium) (1)
38. If (CV is Medium) and (HIC is Low) and (SIO is Low), then (RO is Medium) (1)
63. If (CV is High) and (HIC is High) and (SIO is Medium), then (RO is High) (1)
64. If (CV is High) and (HIC is High) and (SIO is High), then (RO is High) (1)

Table 3 provides the obtained results, and the lowest average testing error was obtained using **trimf** membership functions with dataset A.

Table 3. Evaluation of various MF with different data splits

Data	***Trimf***		***Gbellmf***		***Gaussmf***		***Trapmf***	
	Train	Test	Train	Test	Train	Test	Train	Test
A	0.0139	0.0146	0.0177	0.0195	0.0202	0.0217	0.0290	0.0279
B	**0.0143**	**0.0137**	0.0189	0.0197	0.0216	0.0221	0.0455	0.0453
C	0.0143	0.0141	0.0188	0.0198	0.0201	0.0214	0.0451	0.0472

D	0.0144	0.0143	0.0184	0.0252	0.0188	0.0233	0.0276	0.0255

6 Conclusion

Predicting risk assessment is a complex issue because many factors affect grid computing directly and indirectly. In this paper, we used 3 risk factors, and the ANFIS model was selected based on the minimum value of root mean square error, which is constructed using four triangular-shaped membership function for each input variable and linear membership function for output. Hence we have developed a risk prediction model for computational grid environment using ANFIS.

Acknowledgement

This work was supported by the IPROCOM Marie Curie Initial Training Network, funded through the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007–2013/, under REA grant agreement number 316555.

References

[1] S. A. Abdelwahab, A. Abraham, "A Review of the Risk Factors in Computational Grid," *Journal of Information Assurance & Security,* vol. 8, 2013.

[2] S. J. A. Hossain, Nafis, "Adaptive neuro-fuzzy inference system (ANFIS) based surface roughness prediction model for ball-end milling operation," *Journal of Mechanical Engineering Research,* vol. 4, pp. 112-129, 2012.

[3] K. B. B. Bey, Farid Mokhtari, Aicha Guessoum, Zahia, "CPU load prediction model for distributed computing," in *Parallel and Distributed Computing, 2009. ISPDC'09. Eighth International Symposium on*, 2009, pp. 39-45.

[4] B. D. Karthika, Paresh Chandra, "Prediction of Air Temperature by Hybridized Model (Wavelet-ANFIS) Using Wavelet Decomposed Data," *Aquatic Procedia,* vol. 4, pp. 1155-1161, 2015.

[5] K. G. Djemame, Iain Padgett, James Birkenheuer, Georg and M. K. Hovestadt, Odej Voss, Kerstin, "Introducing risk management into the grid," in *e-Science and Grid Computing, 2006. e-Science'06. Second IEEE International Conference on*, 2006, pp. 28-28.

[6] A. D. Sangrasi, Karim, "Component level risk assessment in grids: A probabilistic risk model and experimentation," in *Digital Ecosystems and Technologies Conference (DEST), 2011 Proceedings of the 5th IEEE International Conference on*, 2011, pp. 68-75.

[7] C. Z. Negoita, L Zimmermann, H, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy sets and systems,* vol. 1, pp. 3-28, 1978.

[8] C. F. Carlsson, Robert, "Risk Assessment of SLAs in Grid Computing with Predictive Probabilistic and Possibilistic Models," in *Preferences and Decisions*, ed: Springer, 2010, pp. 11-29.

[9] I. D. Gourlay, Karim Padgett, James, "Reliability and risk in grid resource brokering," in *Digital Ecosystems and Technologies, 2008. DEST 2008. 2nd IEEE International Conference on*, 2008, pp. 437-443.

[10] N. G. Raju, Y Liu Leangsuksun, Chokchai Box Nassar, Raja and S. Scott, "Reliability Analysis in HPC clusters," in *Proceedings of the High Availability and Performance Computing Workshop*, 2006.

[11] S. A. Abdelwahab, A. Abraham, "Data Mining Approach for Modeling Risk Assessment in Computational Grid," in *Computational Intelligence in Data Mining-Volume 3*, ed: Springer, 2015, pp. 673-684.

[12] M. A. S. Hall, Lloyd A, "Feature subset selection: a correlation-based filter approach," 1997.

[13] O. F. W. Rana, Martijn Quillinan, Thomas B Brazier, Frances Cojocarasu, Dana "Managing violations in service level agreements," in *Grid Middleware and Services*, ed: Springer, 2008, pp. 349-358.

[14] R. H. S. Syed, Maxime Bourgeois, Julien, "Protecting grids from cross-domain attacks using security alert sharing mechanisms," *Future Generation Computer Systems,* vol. 29, pp. 536-547, 2013.

[15] A. Chakrabarti, "Taxonomy of Grid Security Issues," in *Grid Computing Security*, ed: Springer, 2007, pp. 33-47.

[16] H. M. C. Lee, Kwang Sik Jin, Sung-Ho Lee, Dae-Won Lee, Won Gyu Jung, Soon Young Yu, Heon Chang, "A fault tolerance service for QoS in grid computing," in *Computational Science—ICCS 2003*, ed: Springer, 2003, pp. 286-296.

[17] M. F. Smith, Thomas Engel, Michael Freisleben, Bernd, "Countering security threats in service-oriented on-demand grid computing using sandboxing and trusted computing techniques," *Journal of Parallel and Distributed Computing,* vol. 66, pp. 1189-1204, 2006a.

[18] A. D. Chakrabarti, Anish Sengupta, Shubhashis, "Grid computing security: A taxonomy," *IEEE Security & Privacy,* vol. 6, pp. 0044-51, 2008.

[19] S. S. Kar, Bibhudatta, "An Anomaly Detection System for DDOS attack in Grid Computing," *International Journal of Computer Applications in Engineering, Technology and Sciences(IJ-CA-ETS),* vol. 1, p. 553, 2009.

[20] S. R. Naqvi, Michel, "Threat Model for grid security services," in *Advances in Grid Computing-EGC 2005*, ed: Springer, 2005, pp. 1048-1055.

[21] A. M. Barua, Lalitha Snigdha Kosheleva, Olga, "Why trapezoidal and triangular membership functions work so well: Towards a theoretical explanation," 2013.