Atta-ur-Rahman et al. Journal of Cloud Computing: Advances, Systems and Applications (2019) 8:17
https://doi.org/10.1186/s13677-019-0144-9

Journal of Cloud Computing: Advances, Systems and Applications

RESEARCH Open Access

A Neuro-fuzzy approach for user behaviour classification and prediction

Atta-ur-Rahman[1], Sujata Dash[2], Ashish Kr. Luhach[3], Naveen Chilamkurti[4], Seungmin Baek[5] and Yunyoung Nam[5*]

Abstract

Big data and cloud computing technology appeared on the scene as new trends due to the rapid growth of social media usage over the last decade. Big data represent the immense volume of complex data that show more details about behaviours, activities, and events that occur around the world. As a result, big data analytics needs to access diverse types of resources within a decreased response time to produce accurate and stable business experimentation that could help make brilliant decisions for organizations in real-time. These developments have spurred a revolutionary transformation in research, inventions, and business marketing. User behaviour analysis for classification and prediction is one of the hottest topics in data science. This type of analysis is performed for several purposes, such as finding users' interests about a product (for marketing, e-commerce, etc.) or toward an event (elections, championships, etc.) and observing suspicious activities (security and privacy) based on their traits over the Internet. In this paper, a neuro-fuzzy approach for the classification and prediction of user behaviour is proposed. A dataset, composed of users' temporal logs containing three types of information, namely, local machine, network and web usage logs, is targeted. To complement the analysis, each user's 360-degree feedback is also utilized. Various rules have been implemented to address the company's policy for determining the precise behaviour of a user, which could be helpful in managerial decisions. For prediction, a Gaussian Radial Basis Function Neural Network (GRBF-NN) is trained based on the example set generated by a Fuzzy Rule Based System (FRBS) and the 360-degree feedback of the user. The results are obtained and compared with other state-of-the-art schemes in the literature, and the scheme is found to be promising in terms of classification as well as prediction accuracy.

Keywords: Behaviour analysis, Classification, Prediction, FRBS, 360-degree feedback, Neuro-fuzzy, Big data, Cloud computing

Introduction

Online user behaviour analysis is an important area of research that enables different characteristics of users to be studied. The behaviour study and prediction of user intention towards certain products help business houses and industries find the target areas to focus on. The prediction of user intention is based on the interactions within a website [1], which is crucial for retargeting. Usually, e-commerce sites and ad display networks keep track of the search patterns of users to understand their intentions and behaviours. In addition, understanding the demand and making relevant information available for online users on the pretext of the explosive growth of information on the Internet enforces analysis and modelling of the web navigation behaviour of web users. Web mining techniques have become handy to analyse and extract useful information [2] from web content, and they are categorized into web content mining, web usage mining and web structure mining. In a nutshell, web documents are extracted by web content mining, user browsing activities are analysed by web usage mining, and the physical link structure of websites is analysed by web structure mining. Weblogs, which contain user website navigation information such as IP address, client/user id, date, time, method, status code and size of the object, are used by web usage data mining to predict user behaviour. The log is an array of user transactions

* Correspondence: ynam@sch.ac.kr
[5]Soonchunhyang University, Asan, South Korea
Full list of author information is available at the end of the article

that is updated [3] every time the user accesses websites. The web log data are then processed by a sequence of operations [4], such as Data Cleaning, User Identification, Session Identification, Enhanced Pattern Tree Construction, and Pattern Recognition, to retrieve the usage pattern of each user of the website based on his/her activities during the browsing, such as clicks, visited regions, revisits, etc. Many soft computing and data mining algorithms have been suggested by researchers to identify useful patterns in the user's web profile.

Dynamic consolidation frameworks typically consist of many overlapped domains [5], which have been divided into five main subsystems, of which the workload prediction subsystem uses a clustering process, VM and user behaviour estimation, prediction window size, and forecasting process. Authors in [6] enhanced the preciseness of the prediction accuracy of the regression models regarding the changing workload patterns by isolating and studying the impact of the risk minimization principle. The increasing popularity and developments of the e-commerce portals lead to the need to address requirements for privacy and security [7] and help to evaluate shopping behaviour in various domains within the context of mobile devices and the cloud [8].

This paper proposes an automated monitoring and prediction tool, particularly for organizations where there are restrictions on web usage or network access, i.e. each user is given certain privileges and is restricted from certain accesses. This is a common scenario in almost every organization around the globe; hence, there is a dire need to observe users' activities to prevent any unwanted event, such as data theft, virus injection, sniffing and spoofing, etc. The neuro-fuzzy-based prediction system monitors the history of network/web usage of users and then predicts the behaviour as one of the predefined categories. Although the network is equipped with a monitoring system that prevents users from performing the restricted tasks, this system focuses on studying the intentions of users who attempt to conduct the restricted tasks from time to time and reveals the proneness of a user to committing intentional mistakes. The fuzzy rule-based system (FRBS) receives three input variables, namely, normalized web frequency, normalized network frequency and normalized machine frequency, and predicts one output variable named "Suspectedness" that represents the user's tendency to attempt something suspicious. To make the scheme effective and robust, a Gaussian Radial Basis Function Neural Network is extensively trained based on the examples duly generated by FRBS. Once the network is sufficiently trained, it can readily classify a user based on the provided input parameter composed of user characteristics.

The remainder of the paper is organized as follows. Section 2 is dedicated to a literature review, and section 3 describes the proposed scheme. In section 4, the proposed scheme is simulated and results are analysed, while section 5 concludes the paper.

Literature review

Predicting the behaviour of web users is an evolving area of web data mining concerning optimisation of web applications based on the study of web user behaviour. Correctly identifying the behaviour of online users is a challenging task because it requires accurate identification of malicious users from the legitimate users. An analysis of the characteristics of online user behaviour models is studied in [9], and the results show that feature extraction techniques, such as principle component analysis (PCA), independent component analysis (ICA) and self-organizing maps (SOM), can be used to correctly detect anomalies in user behaviour. REPTree and neural network are studied with different parameters and found to be the best models in comparison to others. Various behavioural evaluation techniques have been investigated [10], and their appropriateness for E-learner behaviour evaluation has been analysed. The results of the study established that Kernelized Fuzzy C-mean (KFCM) Clustering is a better technique than simple Fuzzy C-Means (FCM) for behaviour evaluation.

Phishing website detection and identification is a complex problem involving many criteria. A multi-layer neural network framework suggested by [11] accurately classifies and predicts phishing websites. In the past few decades, smartphones and tablets have been used to store a plethora of information related to our day-to-day life, which has prompted mobile attackers to develop malicious applications targeting Android machines. F. Martinelli et al. [12] designed a deep learning classifier on a recent dataset to address this issue and experimentally showed the effectiveness of the model by achieving encouraging results. Another investigation is made by [13] based on user behaviour within a large e-commerce site for predicting the buying intention of customers. They proposed a model consisting of Deep Belief Networks and Stacked Denoising auto-Encoders and showed that the extraction of features from high-dimensional data achieves a substantial improvement. The approach used to analyse user behaviour has great potential in optimising web online advertising [14] and ad-serving systems [15].

Machine learning, data mining, statistics, pattern recognition, and graph theory techniques are used to extract useful information from social network sites. However, graph theory [16] is used to detect anomalies in the user behaviour. Similarly, researchers studying the anomalies in user data and abnormalities in user behaviour use PCA [17]. The study of human behaviour characteristics on social media based on their post and reply behaviours at different times of a day [18] on online forums has revealed some interesting features that are

Table 1 Web data sources

#	Log type	Log Format	Example Log	Extracted Information
1	Web Server Logs	•W3C Extended Log File Format •Microsoft IIS Log Format •NCSA Common Log File Format •ODBC Logging	#Fields: time c-ip cs-method cs-uri-stem sc-status cs-version 18:22:15173.18.255.255 GET /default.htm 200 HTTP/1.0	• On May 8, 1999 at 6:22 P.M. UTC • A user with HTTP version 1.0 and the IP address of 173.18.255.255 issued an HTTP GET command for the file Default.htm • The request was returned without error.
2	Microsoft IIS Log Format	• The user's IP address • User name • Request date and time • HTTP status code • The number of bytes received. Also: • The elapsed time of the request • The number of bytes sent • The action (for example, a download carried out by a GET command) and the target file. • The items are separated by commas • The time is recorded as local time.	192.168.114.201,—,03/20/98,7:55:20,W3SVC2, SALES1, 192.168.114.201, 4502,163,3223,200,0,GET,DeptLogo.gif 172.16.255.255,anonymous, 03/20/98,23:58:11, MSFTPSVC, SALES1,192.168.114.201, 60,275,0,0,0, PASS,intro.htm	• An anonymous user with the IP address of 192.168.114.201 issued an HTTP GET command for the image file DeptLogo.gif at 7:55 A.M. on March 20, 1998 From a server named SALES1 at IP address 172.21.13.45. • The 163-byte HTTP request had an elapsed processing time of 4502 milliseconds (4.5 s) to complete, and returned, without error, 3223 bytes of data to the anonymous user.
3	NCSA Common Log File Format	• Remote host name • User name • Date • Time • Request type • HTTP status code • Number of bytes received by the server.	172.21.13.45 — REDMOND\fred [08/Apr/1997:17:39:04–0800] "GET /scripts/iisadmin /ism.dll?http/serv HTTP/1.0" 2,003,401	A user named Fred in the REDMOND domain, with the IP address of 173.21.13.45, issued an HTTP GET command (that is, downloaded a file) at 5:39 P.M. on April 8, 1998. The request returned, without error, 3401 bytes of data to the user named Fred.

helpful for detecting spam and spammers. The understanding of user behaviour in online-social-network (OSN) is a great challenge for social network analysis. The study and analysis of the behaviour of the users who are members of two social networks found important specificities [19] about their privacy settings, their selection of friends and the activities they perform, and more specifically, the analysis is consistent with the recent findings. In addition, many algorithms, such as Fuzzy Rule Based System (FRBS) [20], Association Rule Mining, Linear Regression [21], REPTree [22], etc., are employed to classify user behaviour depending on their past web usage activities, and in turn, it helps in maintaining network security and privacy. In [23], the authors summarize automated opinion mining from web usage logs and social networks, including Twitter, Facebook, etc. It is further narrated that various techniques have been utilized in this regard, including machine learning, deep learning, text and data mining, etc. It is concluded that machine-learning approaches are better in terms of classification accuracy and robustness [24–26].

Behaviour analysis and prediction could be used in many fields of real life, for example, in the educational domain: predicting students' progress, trends in a particular subject, evaluation of a course/teacher [27–30], and dropout rate in an institute [31]. Researchers also study classification and prediction of customer shopping trends concerning a particular product [32], which is of prime importance in e-business and e-commerce. In [33], the authors investigated a recurrent neural network for human behaviour prediction. The purpose was to provide good governance to people having risks related to slight mental impairment and weakness. A long-short term memory network model was developed by calculating actions, activities and inter-activity and intra-activity behaviours. Consequently, the model could predict the next possible behaviour of the person. The techniques [34, 35] provide human behaviour analysis for the sake of predicting the next possible action based on his/her previous traits and activities performed and observed by various techniques, such as feedback, monitoring sensors, etc., in a particular environment. In [34], the researchers combined data-driven and knowledge-driven techniques for activity modelling. A clustering algorithm was developed to extract knowledge from a domain expert's knowledgebase. Consequently, the technique has been tested with real user input, noisy sensors and challenging activity arrangements,

Table 2 Machine and Network Data sources

#	Log Type	Information Contained in the Log
1	ODBC Logging	• The user's IP address • User name • Request date and time • HTTP status code • Bytes received • Bytes sent • Action carried out (for example, a download carried out by a GET command) • The target (for example, the file that was downloaded). • The time is recorded as local time • You must specify the database to be logged to. • You must setup the database table manually to receive the data.
2	Proxy Server Logs	• Apache with mod_proxy • Apache Traffic Server • HAProxy • Internet Information Services configured as proxy module • Nginx • Privoxy • Squid • Varnish which is reverse proxy only • WinGate
3	Browser History	All web Internet travellers sustain the user visit history in one structure or another. The web browser makes a .dat data file, while Chrome keeps the data in multiple places. SQLite keeps a data file in their specific data files on the system drive. These data files can be analysed at playback to draw out useful information for web customization purposes.
4	Server or Client-side Visit Logger App	One can develop small applications to run on the server or customer side to collect information and location and check out history information. Usually, such applications are launched on the customer side as add-ons for the web browser or integrated inside the website rules; a widely used example of such an app is search engines analytics, where a small piece of a rule provided by search engines is placed in the footer of each web page. That rule gathers customer geographical info, and the web page checks out the details, including the names of pages visited, customer recommendation website (from which website did the customer come from before arriving at the current website), and in the case of recommendation, also gathers any keyword search; it also maintains customer stay time on a particular web page to obtain an understanding of users' interests and the website bounce rate, i.e. how many customers left the website from the site they landed on attaining the particular website.

and 100% accuracy was obtained, but at the cost of learning some false-positive models. In [35], a similar approach for activity identification was proposed. In [36], researchers provide a technique for identifying daily life activities to assist elderly and mentally challenged persons. The proposed technique outperforms the existing knowledge-driven and ontology-based approaches by employing Latent Dirichlet Allocation (LDA) topic modelling [36] and testing over Kasteren [37] and Ordonez datasets [38]. In short, user behaviour classification, modelling, profiling and prediction for the sake of different benefits in different fields of life, such as trend analysis, security and privacy, e-commerce, education, banking, medicine, etc., is the hottest area of research in data mining and machine learning [39]. Because it is a complex phenomenon, the effectiveness of schemes may vary over the domains. Thus, to help address this issue, hybrid intelligent techniques composed of evolutionary and soft computing techniques [40, 44] have been investigated in this research.

Although numerous approaches have been designed for user classification in the literature, their application areas are too generic, and the domain is wide. Moreover, their main interest was to find a user's trait relative to an entity, a product or an activity. In this research, a neuro-fuzzy-based customized user monitoring system to continuously monitor users' activities within an organization by augmenting his/her 360-degree feedback is proposed, where the rules regarding user behaviour are set by the organization.

Proposed scheme

The system model considered for the research is an organization where there are several departments connected through a local area network (LAN). The employees/users are allowed/disallowed to perform certain activities on their machines, the LAN and the web. For example, as far as machine restrictions are concerned, users are not allowed to insert any flash drive because of the organization's policy. Even the USB ports are disabled by the administrator; however, any attempt in this regard is recorded and logged. Similarly, as far as the network is concerned, users are only allowed to visit their privileged areas that vary from user to user based on his/her role. An admin has one role, a manager has a different role, etc. Any attempt to reach a restricted area is prohibited yet recorded in the form of logs. There are many servers on the network, such as database servers, web servers, application servers, proxy servers, etc., where the logs are maintained. Likewise, the web usage is also restricted, and users are not allowed to browse certain websites. For example, only the organizational email server is allowed, where all of the emails are scanned. Other email servers, such as Gmail, Yahoo, etc., are not allowed. Similarly, there are rules for access to certain websites, and so on. This is a common scenario in almost every organization around the globe; hence, there is a dire need to observe users' activities to prevent any unwanted event, such as data theft, virus injection, sniffing and spoofing, etc.

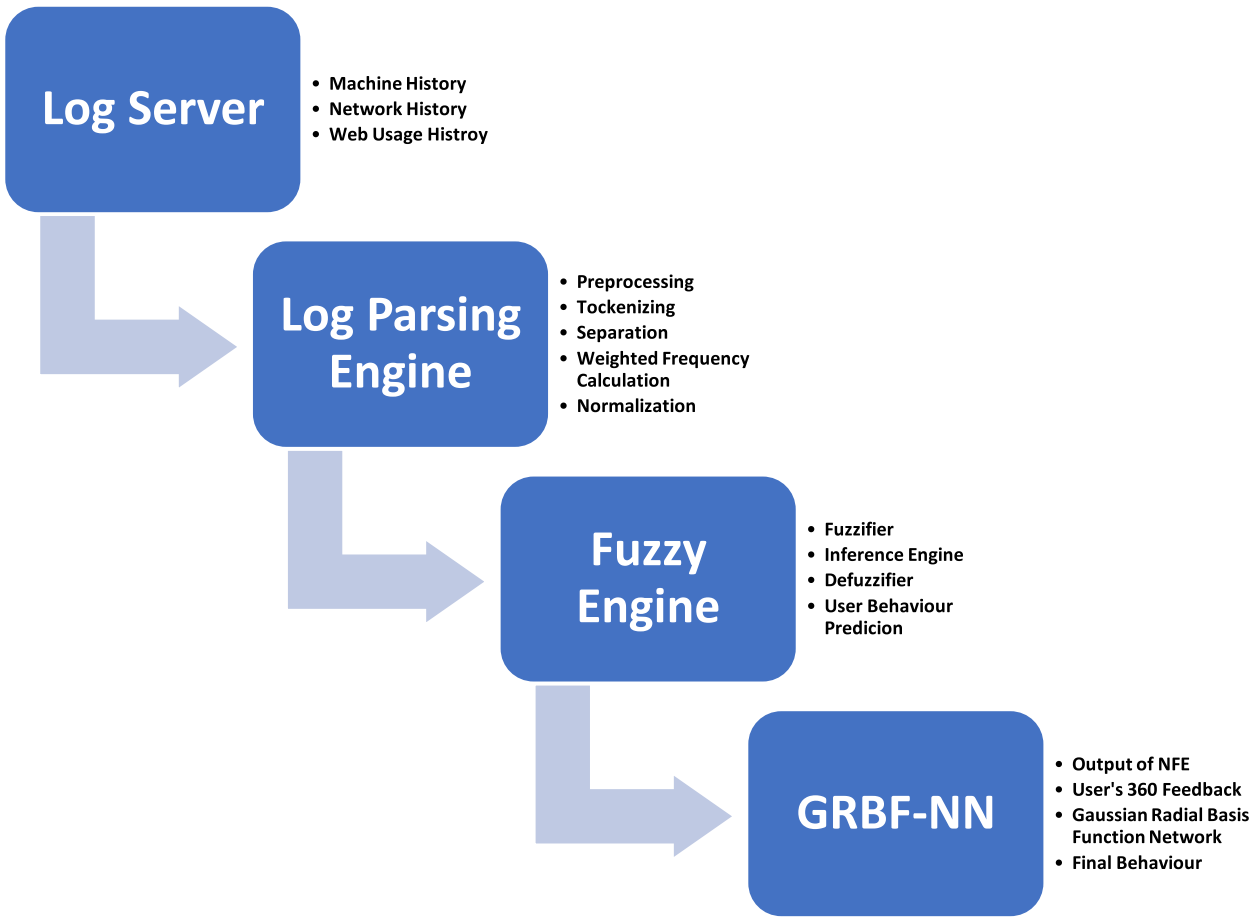


Fig. 1 Schematic of the proposed system model

Data sources

The machine, network and web utilization information primarily preserve records of accessibility styles of the user/visitors. This information can also include the user's visibilities, bookmarks, cookies, modification information, customer concerns and any other communications of the consumer while on the web page. For easy manageability and convenience, the information is arranged into three sections: System Variety Logs, Entrance Variety Logs and Client Web browser Logs. The web server preserves crucial details for network utilized excavation,

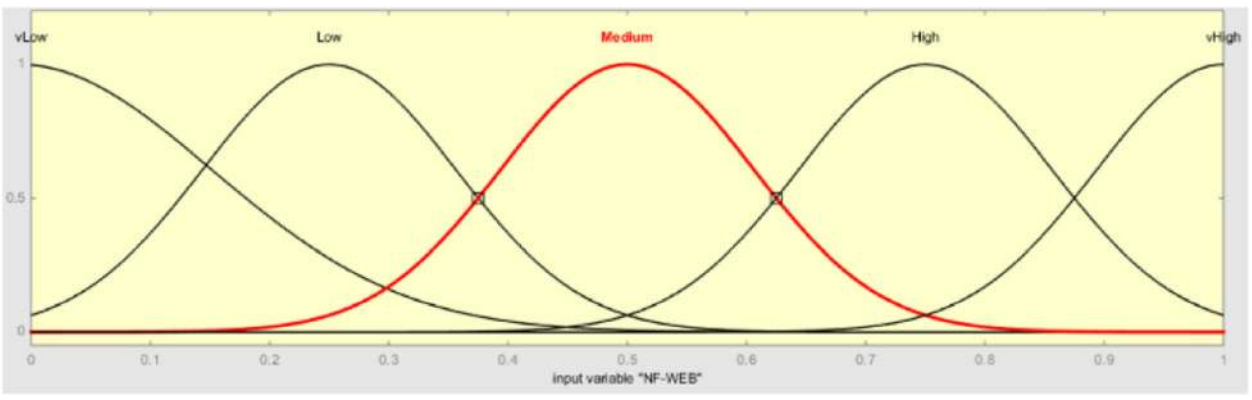


Fig. 2 First input variable "Normalized Web Frequency"

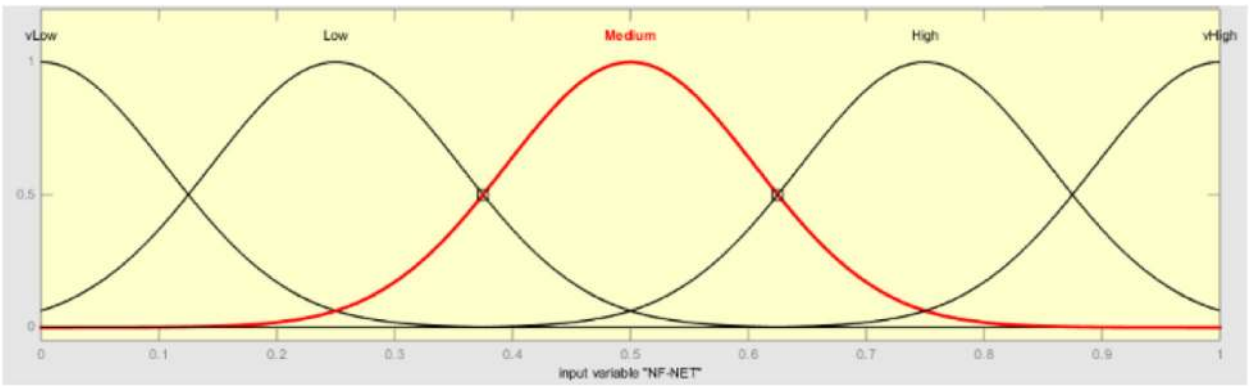


Fig. 3 Second input variable "Normalized Network Frequency"

these records are in general accessibility of websites by multiple users. Each of the records contain the IP address of the user, request time, Consistent Resource Locator, HTTP status cipher, etc. Of course, the details collected are in several standard types, such as log information structure, expanded log information structure, etc., is a portion of a network organized log in W3C. Information for the machine, network and web exploration can be gathered from three sources described below. Two of them are of almost the same type, "Web server logs" and "Proxy server logs," while the third one has different characteristics and framework than the other two sources. Details regarding all three are given in Table 1 and Table 2. These logs are duly collected for one calendar year, from an organizational centralized network server, for the sake of this study.

The information gathered in this process comes from three different locations:

- Server-side collection - the browser behaviour of the web user is gathered in the log file of the site server.
- Client-side collection - uses a user-side application, such as a remote agent, to gather the information of the customer navigation.
- Web proxy server-Web proxy's server takes HTTP demand from the user, the user sends demand to the web server via proxy servers. Proxy-server development is a trial. Advanced network development, such as TCP/IP, is required for this development. The demand interception is limited.

LOG data processing

The overall information planning process is briefly described in the following segments described in Fig. 1. First, a log monitoring server is dedicated to collecting all types of user usage logs, e.g. machine, network and web usage. The logs are in the form of events driven from the Windows Server, which comprises all the types given in Tables 1 and 2 for web, machine and network logs. Due to heterogeneity, these logs need careful parsing and are hence fetched by the Log Parsing Engine, which consists of the following steps.

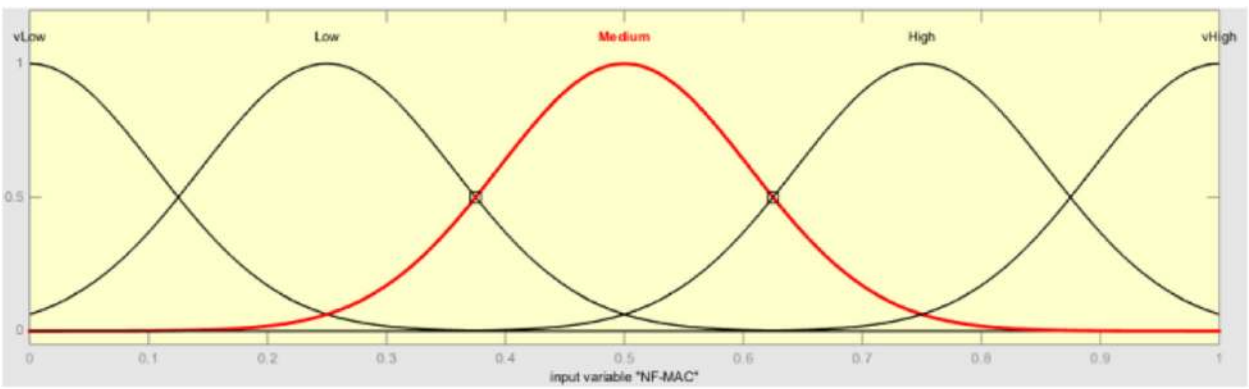


Fig. 4 Third input variable "Normalized Machine Frequency"

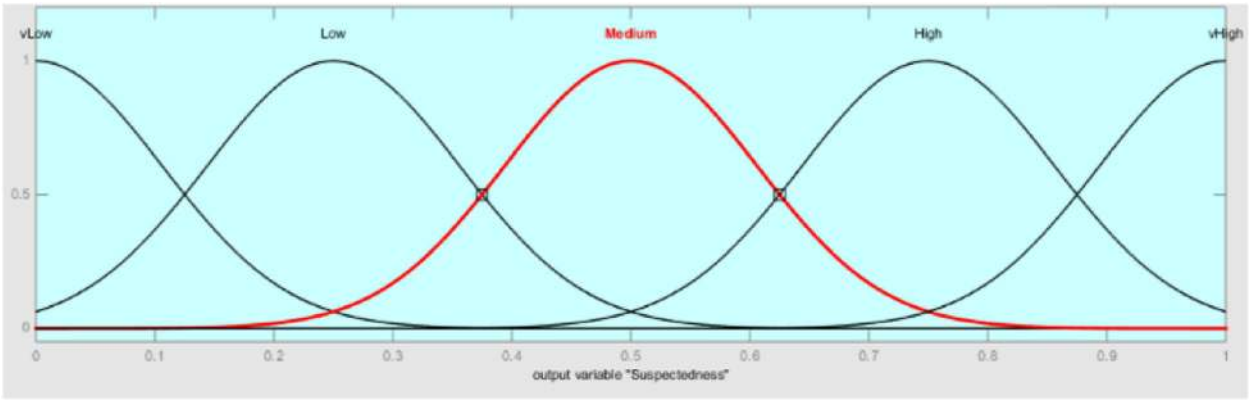


Fig. 5 Output variable "Suspectedness"

Log parsing engine

Not all of the entries of the user's log are useful for the sake of analysis. Thus, irrelevant information must be discarded prior to further data analysis. For example, accesses to unrelated products (such as key images), accesses by Web spiders (i.e. non-human accesses), and unsuccessful demands should be eliminated.

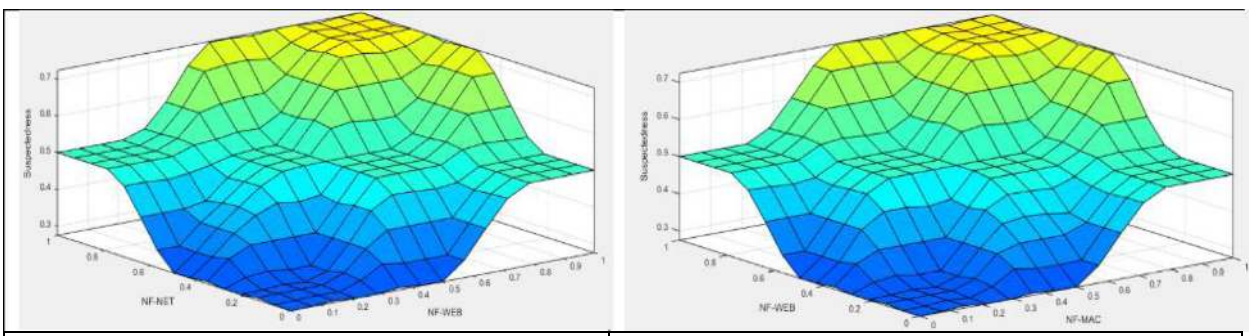
Weighted frequency calculation The log files are in the form of text files. This block removes the noise words and unnecessary information and calculates the frequencies of each violation type obtained from the respective log, namely, web frequency, network frequency and machine frequency. Frequency is an important factor, based on which the gravity of the violation may vary. Similarly, the second factor associated is the weight of the violation. Mathematically,

$$F_{Web} = \sum_{m=0}^{M-1} \omega_m f_m \quad (1)$$

Here, M is the total number of website categories being monitored, e.g. email, online shopping, safe/unsafe websites, social network sites, entertainment, etc., where ω_m is the associated weight factor for each website category having a value between 0 and 1 (0 represents the least harmful or safe/allowed websites, and 1 represents the most harmful or disallowed websites) and f_m represents the frequency of visit/usage of that type. The categorization and the weight assignment of each category are imposed by the organization and may vary from one organization to another. Similarly, the network log frequencies may be expressed as:

$$F_{Net} = \sum_{n=0}^{N-1} \omega_n f_n \quad (2)$$

Here, N is the total number of network activities being monitored, e.g. FTP, shared folders, user area, etc., where ω_n and f_n are the weight assigned to each network category and the frequency of access to that category, respectively. Similarly, the machine log frequencies may be expressed as:



Suspectedness as a function of network and web activity | Suspectedness as a function of web and machine activity

Fig. 6 Rule surfaces.

Table 3 FRBS parameters

#	Parameter	Value
1	Fuzzifier	Gaussian
2	AND method	MIN
3	OR method	MAX
4	Inference Engine	Mamdani Inference Engine (MIE)
5	De-Fuzzifier	Centre Average Defuzzifier (CAD)
6	Size of the Rule base	5x5x5 = 125
7	Cardinality of FRBS	3 × 1
8	Input/output variable ranges and number of functions	[0–1]; 5

$$F_{Mac} = \sum_{p=0}^{P-1} \omega_p f_p \quad (3)$$

Here, P is the total number of machine activities being monitored, e.g. flash or pen drive attachment, disk I/O, killing processes, etc., where ω_p and f_p are the weight assigned to each machine log category and the frequency of access to that category, respectively.

Normalizing frequencies The frequencies of each type may be a large number that may increase the respective value of the weighted frequency score, so there is a need for normalization. Normalization is performed by dividing the corresponding weighted frequency factor by the number of total categories and the maximum frequency in the respective type. Subsequently, normalization activity will be performed to normalize the frequency of each type. Normalization is subject to the total categories in that type and the maximum frequency in that type. Thus, Eqs. 4–6 represent NF_{Web}, NF_{Net} and NF_{Mac}, namely, normalized web, network and machine frequencies, respectively. These are given by:

$$NF_{Web} = \frac{F_{Web}}{M * max(f_1, f_2, ..., f_M)} \quad (4)$$

$$NF_{Net} = \frac{F_{Net}}{N * max(f_1, f_2, ..., f_N)} \quad (5)$$

$$NF_{Mac} = \frac{F_{Mac}}{P * max(f_1, f_2, ..., f_P)} \quad (6)$$

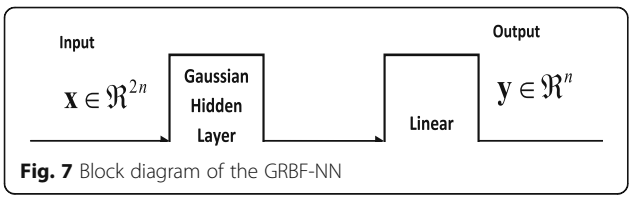


Fig. 7 Block diagram of the GRBF-NN

where M, N and P are the total categories in each type. After normalization, the frequency factor is confined between 0 and 1 so that the inputs become compatible with the input format of the proposed neuro-fuzzy rule-based system.

Design of the Fuzzy Rule-Based System

The design of the fuzzy rule-based system is originally motivated by [21]. There are three input variables to the fuzzy rule-based system (FRBS), namely, normalized web frequency, normalized network frequency and normalized machine frequency. There is one output variable named "Suspectedness" that represents the user's tendency to attempt something suspicious. These variables are shown in Figs. 2, 3, 4 and 5, respectively. The overall rule surfaces are shown in Fig 6. The total number of rules in the rule-based system is the Cartesian product of the number of membership functions in each input variable, so there are one hundred and twenty-five rules in the designed rule base. These rules are formulated based on maximum likelihood (ML) criteria with normal distribution. The design of the fuzzy rule-based system is carried out in the MATLAB Fuzzy Logic Toolbox, and the parameters are given in Table 3. The rule format is given below. IF ((NF-web= "vLow") AND (NF-net= "Medium") AND (NF-mac = "vHigh")) THEN (Suspectedness= "Medium")

User's 360 feedback

The user's 360 feedback is an important measure of an employee's overall behaviour evaluated by the concerned authorities. The common attributes of 360 feedback are usually associated with the employee's performance, contribution, and productivity and are directly linked with his/her appraisal. There are different theories about this feedback, and its pros and cons may vary from organization to organization [41]. In this research, it is considered as a complementing factor to the technological usage behavioural model discussed previously for accurate user profiling.

Gaussian radial basis function neural network (GRBF-NN)

Gaussian Radial Basis Function Neural Networks (GRBF-NNs) are considered the most powerful networks for dynamic and nonlinear systems [42]. The fast, linear learning algorithm is capable of representing complex non-linear mapping and also improves the generalization capability of the network. In this research, GRBF-NNs are suitable due to the dynamic nature of the problem where we need to precisely predict the user's behaviour based on his machine, network and web usage history, as well as his 360-degree feedback. The schematic diagram for the NN is given in Fig. 7.

The following are the steps involved in construction of the network.

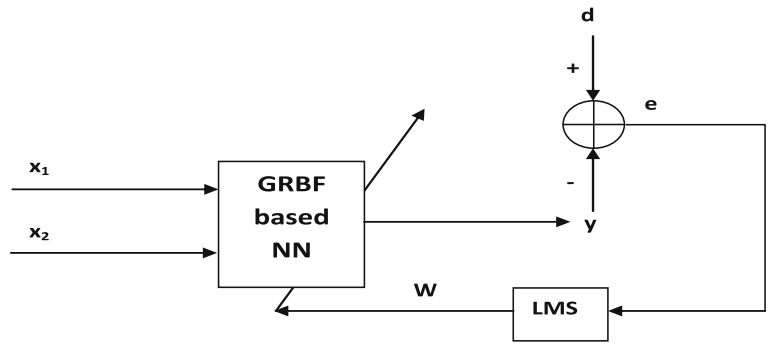


Fig. 8 Network training using the LMS algorithm

Example set

To train the network, a large number of examples are generated by the proposed Fuzzy Rule-Based System and augmented with the user's 360 feedback, where both are in the form of a number. This training is performed in the following manner.

- Choosing a user's behaviour (Suspectedness) from FRBS (× 1)
- Selecting the 360-degree feedback of the same user (× 2)
- Finding the aggregate of both by the following formula:

$$Final\ behaviour(d) = \alpha(360\text{–}degree\ feedback) + (1\text{–}\alpha)(FRBS\ feedback) \tag{7}$$

Here, α ranges between 0 and 1 and corresponds to the factor chosen by an organization that what weight should be assigned to each type of feedback. For example, if the 360-feedback is given a value of 0.7 (70%), the FRBS feedback will be assigned 0.3 (30%), and so on. The FRBS feedback becomes an example of what should be the final behaviour, provided the FRBS feedback and 360-degree feedback. $[\boldsymbol{x}_i^1, \boldsymbol{x}_i^2; \boldsymbol{d}_i]$.

Here, d represents the final behaviour of the user, where lower values represent good behaviour and higher values correspond to bad behaviour.

- A number of examples are generated to make an example set of N in total, i.e. $[\boldsymbol{x}_i^1, \boldsymbol{x}_i^2; \boldsymbol{d}_i]_{i=1}^N$

Training

The network is trained by the supervised Least Mean Square (LMS) algorithm [43]. The training

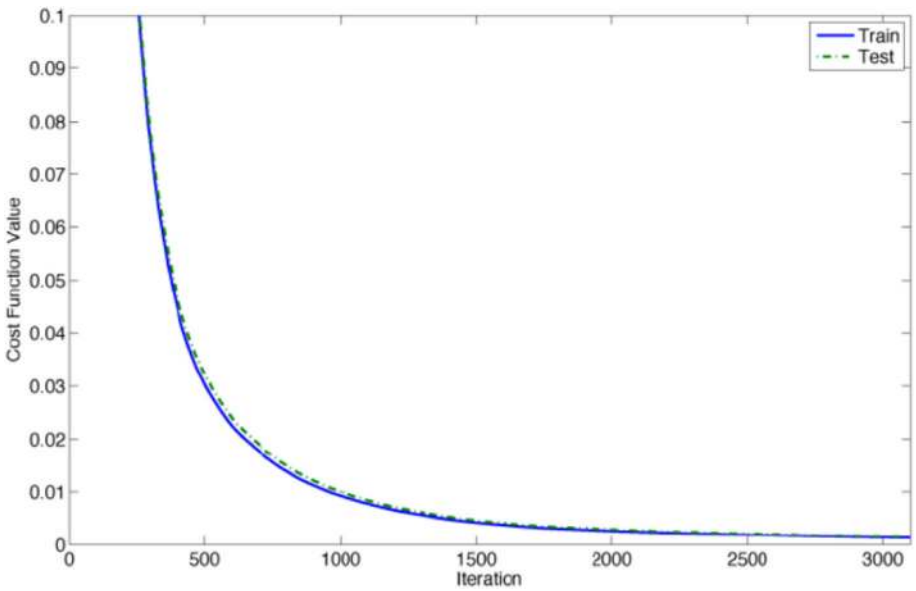


Fig. 9 Convergence Rate

Table 4 users' normalized frequencies

	1	2	3	4	5	6	7	8	9	10	11	12
1	0.814724	0.913376	0.278498	0.964889	0.957167	0.141886	0.792207	0.035712	0.678735	0.392227	0.706046	0.046171
	0.905792	0.632359	0.546882	0.157613	0.485376	0.421761	0.959492	0.849129	0.75774	0.655478	0.031833	0.097132
	0.126987	0.09754	0.957507	0.970593	0.80028	0.915736	0.655741	0.933993	0.743132	0.171187	0.276923	0.823458
	0.624222	0.599827	0.541084	0.699685	0.716811	0.423856	0.738755	0.593821	0.719408	0.407598	0.292691	0.330069
2	0.694829	0.034446	0.765517	0.489764	0.709365	0.679703	0.118998	0.340386	0.751267	0.699077	0.547216	0.257508
	0.317099	0.438744	0.7952	0.445586	0.754687	0.655098	0.498364	0.585268	0.255095	0.890903	0.138624	0.840717
	0.950222	0.381558	0.186873	0.646313	0.276025	0.162612	0.959744	0.223812	0.505957	0.959291	0.149294	0.254282
	0.689899	0.306699	0.540684	0.489057	0.522999	0.432145	0.498382	0.340695	0.504522	0.791759	0.314211	0.450124
3	0.814285	0.349984	0.616045	0.830829	0.917194	0.753729	0.075854	0.779167	0.568824	0.337123	0.311215	0.601982
	0.243525	0.196595	0.473289	0.585264	0.285839	0.380446	0.05395	0.934011	0.469391	0.162182	0.528533	0.262971
	0.929264	0.251084	0.35166	0.549724	0.7572	0.567822	0.530798	0.129906	0.011902	0.794285	0.165649	0.654079
	0.67812	0.280629	0.470149	0.578082	0.666846	0.57339	0.292827	0.627537	0.320179	0.419245	0.317502	0.513679
4	0.689215	0.083821	0.152378	0.996135	0.106653	0.77491	0.084436	0.800068	0.181847	0.136069	0.54986	0.622055
	0.748152	0.228977	0.825817	0.078176	0.961898	0.817303	0.399783	0.431414	0.263803	0.869292	0.144955	0.350952
	0.450542	0.913337	0.538342	0.442678	0.004634	0.868695	0.25987	0.910648	0.145539	0.579705	0.853031	0.51325
	0.68669	0.419009	0.538648	0.481568	0.348069	0.748382	0.285794	0.675813	0.238141	0.584501	0.551383	0.515582
5	0.401808	0.123319	0.417267	0.944787	0.337719	0.111203	0.241691	0.131973	0.575209	0.353159	0.043024	0.731722
	0.075967	0.183908	0.049654	0.490864	0.900054	0.780252	0.403912	0.942051	0.05978	0.821194	0.16899	0.647746
	0.239916	0.239953	0.902716	0.489253	0.369247	0.389739	0.096455	0.956135	0.23478	0.015403	0.649115	0.450924
	0.275181	0.251842	0.420804	0.6912	0.556788	0.394831	0.262984	0.682812	0.314682	0.354071	0.307914	0.646901
6	0.547009	0.188955	0.368485	0.081126	0.486792	0.306349	0.817628	0.378609	0.350727	0.550156	0.207742	0.230488
	0.296321	0.686775	0.625619	0.929386	0.435859	0.508509	0.794831	0.81158	0.939002	0.622475	0.301246	0.844309
	0.744693	0.183511	0.780227	0.775713	0.446784	0.510772	0.644318	0.532826	0.875943	0.587045	0.470923	0.194764
	0.540611	0.404034	0.608066	0.573611	0.452277	0.479588	0.739186	0.561907	0.669502	0.592471	0.305734	0.417258
7	0.225922	0.435699	0.430207	0.979748	0.258065	0.262212	0.221747	0.318778	0.085516	0.02922	0.488609	0.458849
	0.170708	0.311102	0.184816	0.43887	0.40872	0.602843	0.117418	0.424167	0.262482	0.928854	0.578525	0.963089
	0.227664	0.92338	0.904881	0.111119	0.594896	0.711216	0.296676	0.507858	0.801015	0.730331	0.237284	0.546806
	0.257998	0.553324	0.436695	0.467751	0.41452	0.481502	0.250463	0.431548	0.331267	0.536591	0.487755	0.69614
8	0.521136	0.62406	0.367437	0.885168	0.098712	0.679728	0.106762	0.779052	0.890923	0.19781	0.500022	0.609867
	0.231594	0.679136	0.987982	0.913287	0.261871	0.136553	0.653757	0.715037	0.334163	0.030541	0.479922	0.617666
	0.488898	0.395515	0.037739	0.796184	0.335357	0.721227	0.494174	0.903721	0.698746	0.744074	0.904722	0.859442
	0.483792	0.62068	0.467948	0.786977	0.284706	0.468362	0.410267	0.733356	0.639496	0.292362	0.65255	0.613475
9	0.805489	0.239932	0.489901	0.712694	0.059619	0.071445	0.818149	0.149865	0.972975	0.453798	0.08347	0.390938
	0.576722	0.886512	0.167927	0.500472	0.681972	0.52165	0.817547	0.659605	0.648991	0.432392	0.133171	0.83138
	0.182922	0.028674	0.978681	0.471088	0.042431	0.09673	0.72244	0.518595	0.800331	0.825314	0.173389	0.803364
	0.551189	0.390651	0.490113	0.500423	0.3014	0.292276	0.748798	0.413608	0.741621	0.565791	0.226753	0.637705
10	0.060471	0.416799	0.291984	0.984064	0.37241	0.339493	0.052677	0.422836	0.417744	0.701099	0.698106	0.128014
	0.399258	0.65686	0.431651	0.167168	0.198118	0.95163	0.737858	0.547871	0.983052	0.666339	0.666528	0.99908
	0.526876	0.627973	0.015487	0.106216	0.489688	0.920332	0.269119	0.942737	0.301455	0.539126	0.178132	0.171121
	0.321175	0.615645	0.281061	0.416115	0.375345	0.690881	0.306073	0.665871	0.544226	0.666013	0.450986	0.424178

process is shown in Fig. 8. In this process, an example [× 1,× 2] is introduced to the network, the output (y) is compared with the desired output (d), and consequently, the absolute error (e = abs(y-d)) is fed back to the LMS algorithm that updates the weights (w). Then, the change in weights is incorporated in the network.

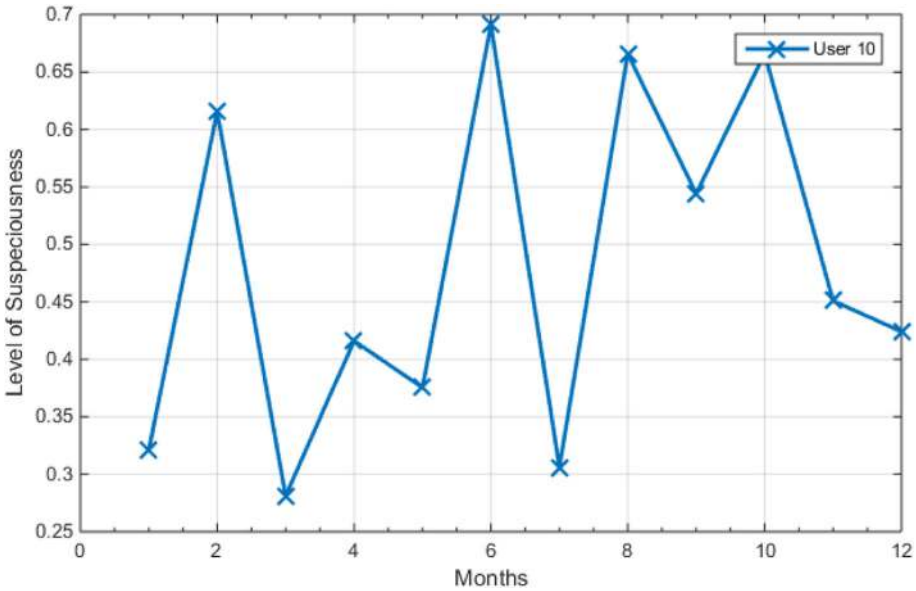


Fig. 10 Single User Behaviour

Simulation results

This section contains the experimental results pertaining to the proposed scheme. In this regard, the dataset is obtained from [21] and contains thousands of pieces of users' data over more than twenty-fine months. The dataset is composed of a mixed log related to machine, network and web usage.

Figure 9 shows the convergence rate of the proposed network during testing and training phases with respect to iterations. This finding shows that, over the iterations, the error rate touches its minimum. The convergence rate goes abruptly after 300 onward iterations and eventually tapers off to zero after 3000 iterations. There is no significant difference

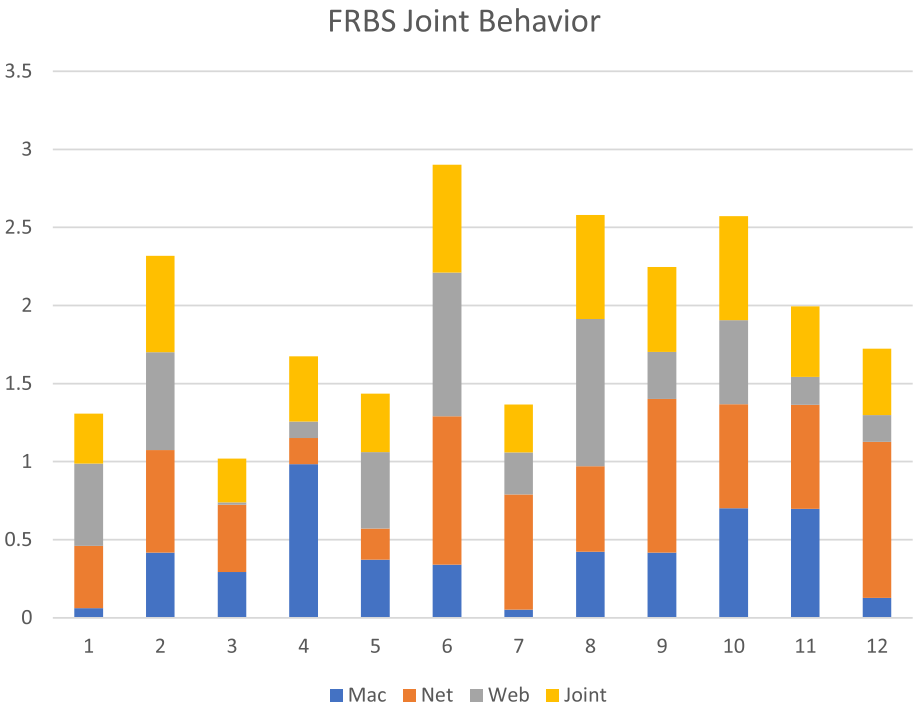


Fig. 11 Joint Behaviour

between the convergences of the testing and training phases mainly because of the fair adoption of examples for each phase.

The sigmoidal function was used as the signature function due to its soft nature of mapping instead of the signum function, which is suitable for hard decision mappings. The normalized machine, network and web frequencies of ten users and the corresponding output of FRBS are enlisted in Table 4 for one calendar year.

Figure 10 shows the behaviour of a single user based on his/her machine, network and web usage. The three inputs are fetched to the FRBS, and the resultant outcome is plotted over the entire year. It can be observed that the user's activities were suspicious in the months of February, June, August and October. During the months of January, March and July, the behaviour was safe, and during April, May, November, and December, it was moderate. The same trend can be observed in the stacked bar graph given in Fig. 11, where the high bars show the most and low bars show the least suspected level, and FRBS output is represented by the colour yellow.

Figure 12 shows the role of 360-degree feedback in conjunction with the proposed FRBS. The combined process is given in Eq. 7, which shows that the weight given to each feedback, that is, 360-degree feedback and FRBS-based feedback, is connected via a variable alpha (α), which is a relative distribution between the two types of feedbacks. If we choose alpha as 70%, then this is the weight of the 360-degree feedback, while the FRBS feedback will have 30%, and so on.

To show the effectiveness of this factor and 360-degree feedback inclusion, three cases are considered in Fig. 12, i.e. the 360-degree feedback values as 70%, 50% and 30%, respectively. The joint behaviour of random users is considered, where the annual average of the FRBS feedback was 0.479797517. The overall behaviour was plotted with respect to alpha ranging between 0 and 1.

At (α = 0), the value is exactly as FRBS; however, with an increase in the value of α, the 360-degree feedback comes into consideration. In this case, the blue bars show the case of 360-degree feedback as 70% (which means users received worse feedback; here, the assumption is that, the higher the number, the worse the feedback, and vice versa); now as α is increasing, the overall behaviour is going to be worse (getting higher). Similarly, when the 360-degree feedback value is 50%, an equal weight is expected for all of the values of behaviour because the FRBS feedback is close to it. At a 360-degree value of 30%, the overall value of the bar is decreasing, which shows that for higher values of α, where the 360-degree feedback received more weight, the overall user's credibility is improved.

In short, the value of α is purely an organizational choice where the administration decides which feedback should be given which weight.

Figure 13a shows the performance of the GRBF Neural Network as a predictor in the case of an individual user. Here, the solid line shows the actual behavioural value; the curved line shows the prediction trend of the Neural Network, while the straight line shows the linear prediction used in [21]. It is apparent from the diagram that the proposed GRBF-NN closely follows the user trend over the monthly data. Because the data are sparse (monthly basis), the network trendline is not very close to the actual value; however, for the dense data, it can

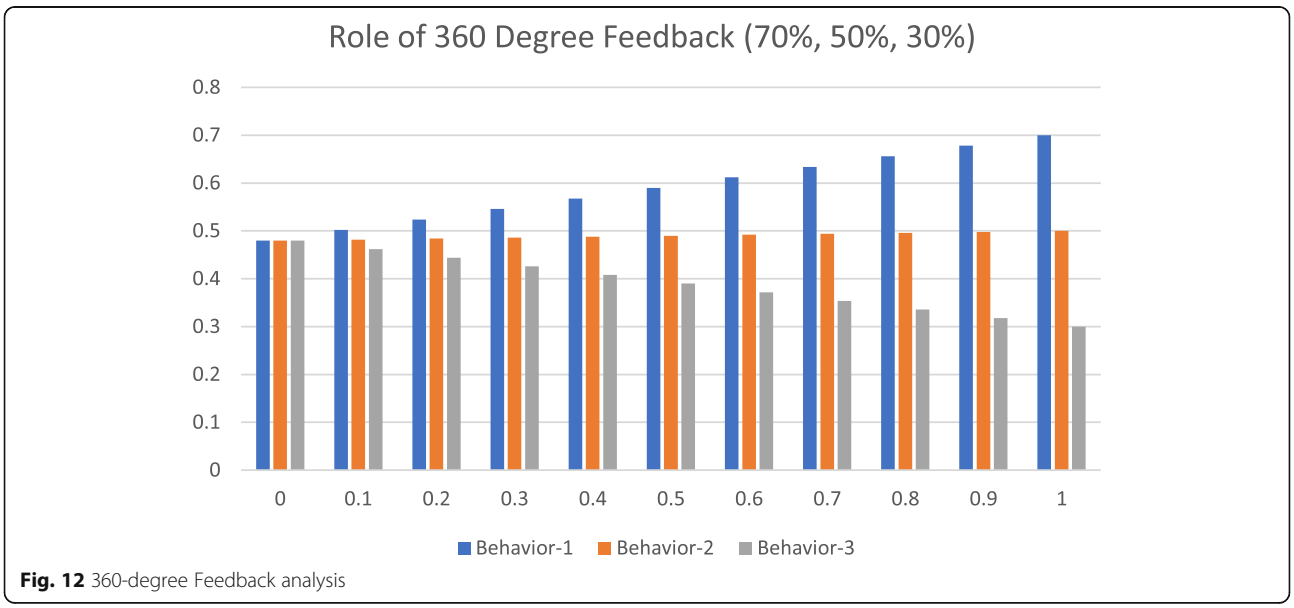


Fig. 12 360-degree Feedback analysis

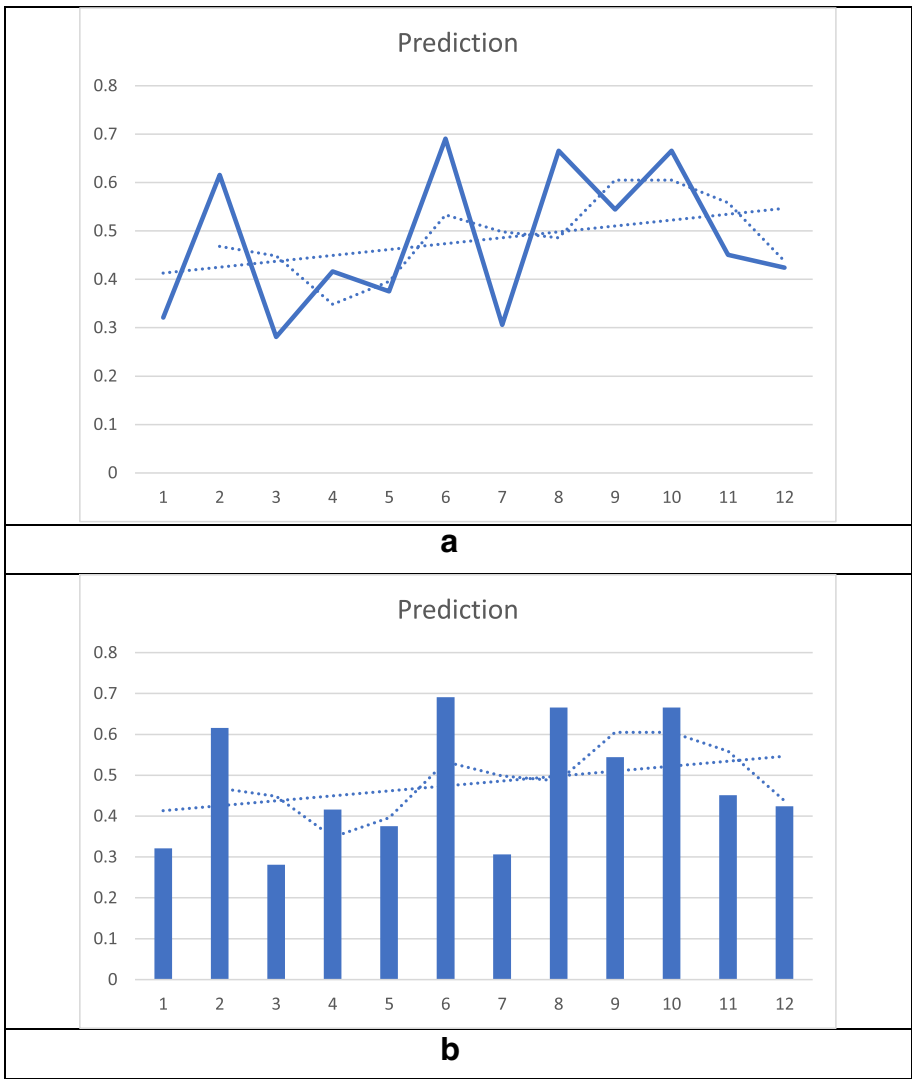


Fig. 13 a GRBF-NN prediction vs. Linear Prediction [21]. Fig. 13b: GRBF-NN prediction vs. Linear Prediction [21]

follow more closely. To further differentiate the performance of the proposed GRBF-NN-based prediction over the linear prediction given in [21], Fig. 13b shows the trend line over the bar chart.

As far as complexity and overhead of the proposed approach is concerned, the technique is composed of three main phases, namely, the dataset generation (using FRBS and 360-degree feedback), training phase and testing phase, respectively. The main complexity is involved in the dataset generation and training phases, which are purely offline processes. Once the network is sufficiently trained, the complexity becomes constant because, as the input appears to the network, it is classified into the corresponding behaviour class regardless of the nature of the example. However, the scheme given in [21] exhibits a relatively higher complexity because, in that case, every time a new example appears at the input, the technique executes the complete classification algorithm to find the final behaviour of the user.

Conclusion

This paper presents a novel technique for user behaviour classification and prediction using a Fuzzy Rule-Based System (FRBS) augmented with 360-degree user organizational feedback (the 360-degree feedback plays a vital role in organizations to precisely classify/ratify an employee) and Gaussian Radial Basis Function Neural Network (GRBF-NN), respectively. The FRBS was designed to classify the user based on his/her machine, network and web usage logs duly collected by a network server of the organization. The logs are pre-

processed in a series of steps, prior to fetching them in FRBS. The designed FRBS and 360-degree feedback are jointly used to produce the example set for the GRBF-NN, which is done by, first, randomly picking examples from the dataset; second, passing through the FRBS, which classifies the user based on his/her logs; and third, by augmenting the same user's 360-degree feedback. Upon sufficiently training the network, it can precisely predict the behaviour of a user on the fly.

In the future, the proposed scheme can be further fine-tuned for further perfection. For example, in the current scheme, while designing FRBS, all types of logs are given equal weights; in the future, the weights may be different for each type of activity depending upon organizational policies. Moreover, the 360-degree feedback is an annual feedback, while the logs are processed on a monthly basis. The alignment can be performed by introducing other types of feedback, such as weekly and monthly feedback. To further fine-tune the results, hybrid intelligent techniques with deep learning concepts may also be investigated.

Abbreviations

FCM: Fuzzy C-Means; FRBS: Fuzzy Rule Based System; GRBF-NN: Gaussian Radial Basis Function Neural Network; ICA: Independent component analysis; KFCM: Kernelized Fuzzy C-mean; LAN: Local area network; LDA: Latent Dirichlet Allocation; LMS: Least Mean Square; ML: Maximum likelihood; OSN: Online-social-network; PCA: Principle component analysis; SOM: Self-organizing maps; VM: Virtual Machine

Acknowledgements

This research was supported by the KIAT (Korea Institute for Advancement of Technology) grant funded by the Korea Government (MOTIE: Ministry of Trade Industry and Energy). (No. N0001791, HRD Program for ICT Convergence Smart Rehabilitation Industrial Education Program) and the Soonchunhyang University Research Fund.

Authors contribution

AuR participated in the design of the study and performed the statistical analysis, SD conceived of the study, and participated in its design and coordination and drafted the manuscript, AKL participated in the dataset preparation and helped to draft the manuscript, NC helped to prepare the experimental model, SB helped to analyze the experimental results and YN helped to develop the overall manuscript and experimental results. All authors read and approved the final manuscript.

Funding

This research was supported by the KIAT (Korea Institute for Advancement of Technology) grant funded by the Korea Government (MOTIE: Ministry of Trade Industry and Energy). (No. N0001791, HRD Program for ICT Convergence Smart Rehabilitation Industrial Education Program) and the Soonchunhyang University Research Fund.

Availability of data and materials

Information for the machine, network and web exploration has been gathered from three sources. Two of which are of almost of same kind "Web server logs" and "Proxy server logs", while the third one is of different characteristics and framework from other two sources. Details regarding all three are given in Table 1 and Table 2 in the text. The information gathered to this process come from three different locations: Server-side collection - the browser behavior of the web user is gathered in the log file of the site server. Client-side collection - uses a user part application, like a remote agent to gather the information, of the customer navigation. Web proxy server-Web proxy's server takes HTTP demand from user, user send demand to web server via proxy servers. Proxy-server development is a trial. Advanced network development, such as TCP/IP, is required for this development. The demand interception is limited.

Competing interests

The authors declare that there are no competing interests regarding the publication of this paper.

Author details

[1]Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia. [2]North Orissa University, Baripada, Odisha, India. [3]The PNG University of Technology, Lae, Papua New Guinea. [4]La Trobe University, Melbourne, Australia. [5]Soonchunhyang University, Asan, South Korea.

Received: 10 June 2019 Accepted: 8 November 2019
Published online: 21 November 2019

References

1. Curme C, Preis T, Stanley HE, Moat HS (2014) Quantifying the semantics of search behavior before stock market moves. Proc Natl Acad Sci 111(32): 11600–11605
2. Pawel W., Mieczyslaw O., & Michal P. (2012). Web user navigation patterns discovery from WWW server log files", IEEE
3. Yan X, Han J, Afsha R (2003) CloSpan: mining: closed sequential patterns in large datasets, proceedings of the 2003 SIAM international conference on data mining, 166–177
4. Saleh M, SHETTY P, Nisha (2018) Analysis of Web Server Logs to Understand Internet User Behavior and Develop Digital Marketing Strategies. Int J Eng Technol 7(4.41):15–21
5. Ismaeel S, Karim R, Miri A (2018) Proactive dynamic virtual-machine consolidation for energy conservation in cloud data centers. J Cloud Comput 7(10):2018
6. Nikravesh AY, Ajila SA, Lung C-H (2017) An autonomic prediction suite for cloud resource provisioning. J Cloud Comput 6(3):2017
7. Shirazi F, Iqbal A (2017) Community clouds within M-commerce: a privacy by design perspective, Community clouds within M-commerce: a privacy by design perspective. J Cloud Comput 6:22
8. Ahmad A, Khan M, Jabbar S, Rathore MMU, Chilamkurti N, Min-Allah N (2017) Energy efficient hierarchical resource management for mobile cloud computing. IEEE Trans Sustainable Comput 2(2):100–112
9. Deshpande D, Deshpande S (2017) Analysis of various characteristics of online user behavior models. Int J Comput Appl 161(11):5–10
10. Hogo MA (2010) Evaluaion of E-learners Behariour using different fuzzy clustering models: a comparative study. Int J Comput Sci Info Secur 7(2): 131–140
11. Martin A, Anutthamaa NB, Sathyavathy M, Manjari M, Francois S, Venkatesan P (2011) A framework for predicting phishing websites using neural networks. Int J Comput Sci Issues 8(2):330–336
12. Martinelli F, Marulli F, Mercaldo F (2017) Evaluating convolutional neural network for effective Mobile malware detection. Procedia Comput Sci 112: 2372–2381
13. Vieira A (2016) Predicting online user behavior using deep learning algorithms, arXiv:1511.06247v1 [cs.LG]; Cornell University, vol. abs./1511. 06247
14. Smith S (2008) Behavioral targeting could change the game. January 23: 2007 Available at: http://www.econtentmag.com/Articles/ArticleReader.aspx?ArticleID=18964
15. Jaworska J,Sydow M (2008). Behavioral Targeting in On-line Advertising: An Empirical Study. In 9th International Conference on Web Information Systems Engineering (Wise 2008), Auckland, 62-76, Auckland, New Zealand, Springer-Verlag Berlin, Heidelberg Springer
16. Bindu P, Thilagam PS (2016) Mining social networks for anomalies: methods and challenges. J Netw Comput Appl 68:213–229
17. Meng B, Jian X, Wang M, Zhou F (2016) Anomaly detection model of user behavior based on principal component analysis. J Ambient Intell Humaniz Comput 7(4):547–554
18. Chen G, Wang N, Zhang F, Jiang H (2015) Understanding the time characteristic of user behavior on online forums. In: In 2015 IEEE international conference on big data, Santa Clara, pp 2300–2306. https://doi.org/10.1109/BigData.2015.7364019

19. Buccafurri F, Lax G, Nicolazzo S, Nocera A (2015) Comparing twitter and Facebook user behavior: privacy and other aspects, computers in human behavior, 10, 87–95. Elsevier Science Publishers B. V. Amsterdam, The Netherlands
20. Rahman AR, Zaidi DN, Salam MH, Jamil S (2013) User behavior classification using fuzzy rule based system. In: 13th international conference on hybrid intelligent systems (HIS'13), 118-123, Tunis, Tunisia
21. Rahman AR, Saleh AA, Abraham A (2017) User behavior classification and prediction using fuzzy rule based system and linear regression. J Inf Assur Secur 11(2017):086–093
22. Haizan WN, Mohamed W, Najib M, Salleh M, Omar AH (2012) A comparative study of reduced error pruning method in decision tree algorithms. In: IEEE international conference on control system, computing and engineering, pp 34-38, Penang, Malaysia
23. Khan SN, Nawi NM, Imrona M, Shahzad A, Ullah A, Rahman AR (2018) Opinion mining summarization and automation process: a survey, *International Journal on Advanced Science*. Eng Inf Technol 8(5):1836–1844
24. Dash, S., Tripathy, B.K., & Rahman, A. R. (Editors), (2017). Modeling, analysis, and applications of nature-inspired Metaheuristic algorithms, ISBN-10: 1522528571 Publisher: IGI Global, USA
25. Fang H, Lu W, Wu F, Zhang Y, Shang X, Shao J, Zhuang Y (2015) Topic aspect-oriented summarization via group selection. Neurocomputing 149: 1613–1619
26. Heu JU, Qasim I, Lee DH (2015) FoDoSu: multi-document summarization exploiting semantic analysis based on social folksonomy. Inf Process Manag 51(1):212–225
27. Rahman AR (2013) Teacher assessment and profiling using fuzzy rule based system and Apriori algorithm. Int J Comput Appl 65(5):22–28
28. Rahman AR, Dash S (2017) Big data analysis for teacher recommendation using data mining techniques. Int J Control Theory Appl 10(18):95–105
29. Rahman AR, Dash S (2017) Data Mining for Students' trends analysis using Apriori algorithm. Int J Control Theory Appl 10(18):107–115
30. Rahman AR, Sultan K, Aldhafferi N, Alqahtani A (2018) Educational data mining for enhanced teaching and learning. J Theor Appl Inf Technol 96(14):4417–4427
31. Haiyang L, Wang Z, Benachour P, Tubman P (2018) A Time Series Classification Method for Behaviour-Based Dropout Prediction. In: 2018 IEEE 18th international conference on advanced learning technologies (ICALT), 191-195, Mumbai. India, https://doi.org/10.1109/ICALT.2018.00052
32. Raju SS, Dhandayudam P (2018) Prediction of customer behaviour analysis using classification algorithms, in AIP conference proceedings 1952. https://doi.org/10.1063/1.5032060
33. Almeida A, Azkune G (2018) Predicting human behaviour with recurrent neural networks. MDPI- Appl Sci 8(305):1–13
34. Azkune G, Almeida A, López-de-Ipiña D L, Chen L (2015) Extending knowledge-driven activity models through data-driven learning techniques. Expert Syst Appl 42:3115–3128
35. Azkune G, Almeida A, López-de-Ipiña D, Chen L (2015) Combining users' activity survey and simulators to evaluate human activity recognition systems. Sensors 15:8192–8213
36. Ihianle IK, Naeem U, Tawil A (2016) Recognition of activities of daily living from topic model. Procedia Comput Sci 98:24–31
37. Van Kasteren T, Englebienne G, Krose BJA (2011) Human activity recognition from wireless sensor network data: benchmark and software. In: Activity Recognition in Pervasive Intelligent Environments. Press, Atlantis, pp 165–186
38. Ordonez J, de Toledo P, Sanchis A (2013) Activity recognition using hybrid generative/discriminative models on home environments using binary sensors. Sensors 13:5460–5477
39. Yousukkee S (2016) Survey of analysis of user behavior in online social network. In: Management and Innovation Technology International Conference (MITicon). https://doi.org/10.1109/MITICON.2016.8025232
40. ZHANG M, WANG Y, CHAI J (2015) Review of user behavior analysis based on big data: method and application. In: In international conference on advances in mechanical engineering and industrial informatics (AMEII 2015), pp 99–103
41. Bracken DW, Rose DS (2011) When does 360-degree feedback create behavior change? And how would we know it when it does? J Bus Psychol 26(2):183–192
42. Gupta M, Jin L, Homma N (2003) Static and dynamic neural networks from fundamental to advance theory. Willey
43. Rahman AR, Qureshi IM, Malik AN, Naseem MT (2014) A real time adaptive resource allocation scheme for OFDM systems using GRBF-neural networks and fuzzy Rule Base system. Int Arab J Inf Technol 11(6):593–601
44. Ihianle IK, Naeem U, Islam S, Tawil AR (2018) A hybrid approach to recognizing activities of daily living from object use in the home environment. Informatics 5(6):1–25

Publisher's Note