How to cite this paper:

Mohamad Farhan Mohamad Mohsin, Azuraliza Abu Bakar, Abdul Razak Hamdan & Mohd Helmy Abdul Wahab (2018). An improved artificial dendrite cell algorithm for abnormal signal detection. *Journal of Information and Communication Technology* (JICT), *17* (1), 33-54.

AN IMPROVED ARTIFICIAL DENDRITE CELL ALGORITHM FOR ABNORMAL SIGNAL DETECTION

[1]Mohamad Farhan Mohamad Mohsin, [2]Azuraliza Abu Bakar, [3]Abdul Razak Hamdan &
[4]Mohd Helmy Abdul Wahab

[1]School of Computing, Universiti Utara Malaysia, Malaysia
[2,3]Faculty of Science & Information Technology Universiti Kebangsaan Malaysia, Malaysia,
[4]Faculty of Electrical and Electronic Engineering Universiti Tun Hussein Onn Malaysia, Malaysia

farhan@uum.edu.my; aab@ukm.my; arh@ftsm.ukm.my; helmy@uthm.edu.my

ABSTRACT

In dendrite cell algorithm (DCA), the abnormality of a data point is determined by comparing the multi-context antigen value (MCAV) with anomaly threshold. The limitation of the existing threshold is that the value needs to be determined before mining based on previous information and the existing MCAV is inefficient when exposed to extreme values. This causes the DCA fails to detect new data points if the pattern has distinct behavior from previous information and affects detection accuracy. This paper proposed an improved anomaly threshold solution for DCA using the statistical cumulative sum (CUSUM) with the aim to improve its detection capability. In the proposed approach, the MCAV were normalized with upper CUSUM and the new anomaly threshold was calculated during run time by considering the acceptance value and min MCAV. From the experiments towards 12 benchmark and two outbreak datasets, the improved DCA is proven to have a better detection result than its previous version in terms of sensitivity, specificity, false detection rate and accuracy.

Keywords: Anomaly threshold, dendrite cell algorithm, multi-context antigen value.

Received: 20 February 2017 Accepted: 30 July 2017

AN IMPROVED ARTIFICIAL DENDRITE CELL ALGORITHM FOR ABNORMAL SIGNAL DETECTION

[1]Mohamad Farhan Mohamad Mohsin, [2]Azuraliza Abu Bakar2, [3]Abdul Razak Hamdan & [4]Mohd Helmy Abdul Wahab

[1]School of Computing, Universiti Utara Malaysia, Malaysia
[2,3]Faculty of Science & Information Technology
Universiti Kebangsaan Malaysia, Malaysia,
[4]Faculty of Electrical and Electronic Engineering
Universiti Tun Hussein Onn Malaysia, Malaysia

farhan@uum.edu.my; aab@ukm.my; arh@ftsm.ukm.my;
helmy@uthm.edu.my

ABSTRACT

In dendrite cell algorithm (DCA), the abnormality of a data point is determined by comparing the multi-context antigen value (MCAV) with anomaly threshold. The limitation of the existing threshold is that the value needs to be determined before mining based on previous information and the existing MCAV is inefficient when exposed to extreme values. This causes the DCA fails to detect new data points if the pattern has distinct behavior from previous information and affects detection accuracy. This paper proposed an improved anomaly threshold solution for DCA using the statistical cumulative sum (CUSUM) with the aim to improve its detection capability. In the proposed approach, the MCAV were normalized with upper CUSUM and the new anomaly threshold was calculated during run time by considering the acceptance value and min MCAV. From the experiments towards 12 benchmark and two outbreak datasets, the improved DCA is proven to have a better detection result than its previous version in terms of sensitivity, specificity, false detection rate and accuracy.

Keywords: Anomaly threshold, dendrite cell algorithm, multi-context antigen value.

Received: 20 February 2017 **Accepted:** 30 July 2017

INTRODUCTION

The dendritic cell algorithm (DCA) is a class of computation intelligence inspired by the principle of human immune systems. Classified as one of the artificial immune system (AIS) algorithms, DCA is modeled after the nature behavior of the human defense system against intruders such bacteria, virus, and parasite based on the concept of the danger theory for use in problem-solving. DCA believes the human immune system is triggered only when a dendritic cell recognizes a danger signal released by an unexpected cell death due to pathogenic infection. The dendrite cell plays an important role as an inspector to recognize pathogens that penetrate the body. Analogized from that task, DCA is modeled to detect anomalies mainly in time series related applications. The preliminary DCA prototype was proposed in 2005 by Greensmith, Aickelin, & Cayzer (2005) into a computer network security system in identifying suspicious network intruders, and then it has been fully implemented as a real-time network intrusion detection system in the following years (Greensmith, Twycross, & Aickelin, 2006). After that, DCA has been seen in various area, mainly to time series anomaly detection-based problems including fault detection (Lee, Lau, Wong, Tam, & Chan, 2016; Ran, Timmis, & Tyrrell, 2010), outbreak detection (Mohamad Mohsin, Hamdan, & Abu Bakar, 2013), and intrusion detection (Anandita, Rosmansyah, Dabarsyah, & Choi, 2015; Bukola & A.O., 2016; El-Alfy & AlHasan, 2016; Ou, 2012). Recently, DCA also has been used as a tool to classify structured and unstructured information (Zainal & Jali, 2017). Their published results exhibit DCA is capable of discovering hidden anomalies well in comparison to other detection systems.

DCA employs the dangers of antigen as a criterion to determine the abnormality of a data point and this strategy makes it differ from other detection algorithms that rely on the pattern-matching approach. In DCA, each data point is viewed as an antigen that is vulnerable to pathogen attacks. During monitoring, DCA tracks antigen health conditions through its life span and accumulates the final score into a variable called multi-context antigen value (MCAV). Acting as a medical profile, MCAV represents the antigen experience in its lifetime based on the frequency of being a mature antigen over total antigen. At the end, the antigen is classified as an anomaly if the MCAV score is greater than the predefined anomaly threshold (Chelly & Elouedi, 2016).

In recent practice, there are three techniques to determine the anomaly threshold. First, is the try and test experiment based on expert recommendation. Second, is the class distribution between abnormal and normal group (Greensmith, 2007), and the last is based on the min MCAV (Song & Qijuan, 2012). The issue with those implementations is that the value needs to be determined before mining based on historical information that causes the new data point to be unrecognizable if the pattern is distinct from the original setting. Besides that, the try and test approach is a time consuming process and highly depending on expert guidance. One of the solutions is by calculating the value in real time during mining. Although the mean MCAV approach is able to skip the pre-determine anomaly threshold, it has a drawback when facing extreme values among MCAV. In this paper, we proposed an adaptive anomaly threshold based on Cumulative Sum (CUSUM) where it involves two folds; determine the new mean MCAV as a threshold and normalizing the MCAV with CUSUM. The improvements were aimed to allow DCA to determine the threshold value during mining and be robust against extreme value such that that it can produce better detection accuracy. The proposed algorithm was compared with the previous DCA with mean MCAV and four evaluation criteria were applied; the sensitivity, specificity, false detection rate and accuracy. In this study, 12 benchmark datasets from several data providers were chosen as experiment data and two outbreak datasets as a case study. The remainder of this paper is organized as follows. It starts by highlighting the dendrite cell algorithm background and discussion on previous works related to MCAV and the anomaly threshold. It is followed by the presentation of the proposed work and the experiment setup. After that, the results and discussion will be presented and finally the concluding remarks.

DENDRITE CELL ALGORITHM

DCA is derived based on the abstraction of the functionality of the danger theory that takes into account our immune system which is activated when a body cell releases a danger signal as response to infection (Matzinger, 2012). Biologically, the main element of the theory, the DCs will recognize the released signals by collecting body cell protein paired with three signals, PAMP, DS and SS, and then monitors their life progress. The monitoring task continues until the cell dies either a 'healthy death' (normal) or 'unhealthy death' (abnormal).

Analogized from the danger theory's mechanism, DCA is formalized into three phases: initialization, updating and aggregation. In the initialization stage, the algorithm parameters are configured and initialized, and all DCs are set in the immature state. During this stage, each item in the dataset is marked as antigen that has chances to be attacked by pathogens. In the updating phase, a continuous process of updating data structures from the input signals and the antigens is performed. The immature DCs collect the input signals

(PAMP, DS, and SS) together with the multiple antigens sampling, calculates the changes and determines which antigen is causing the changes using the accumulative function such that

$$\mathrm{O_j}(\mathrm{x}) = (\textstyle\sum_{i=0}^{i=3} W_{ij} \; * IS_{ij}(x)) \;) \;/\; (\textstyle\sum_{1=0}^{i=3} |\mathrm{W_{ij}}|) \tag{1}$$

where W is the weight matrix, IS is the input signal, OS is the output signal, i represents the PAMP, SS, and DS while j is the output signal categoring CSM, Mature, and Semi-Mature.

All input signals are transformed into three cumulative output signals: CSMs, Mature, and Semi-Mature. Throughout several samplings, the output signals will change the immature DCs[1] state either to semi-mature (normal) or mature (abnormal) depending on the CSM value such that it must be greater than the migration threshold. If the CSM value exceeds the threshold, the type of maturity is determined; 'mature' if the Mature > Semi-Mature or 'semi-mature' if Mature < Semi-Mature.

The aggregation phase occurs when the learning ends. At the final stage, antigens that are presented by the Mature and Semi-Mature context are accessed to determine their abnormalities. Termed as the mature context antigen value (MCAV), the abnormality of an antigen is calculated as

$$\text{MCAV} = (\text{Mature})/(\text{Semi Mature} + \text{Mature}) \tag{2}$$

If the MCAV is above a predetermined value (anomaly threshold), the antigen is labeled as abnormal/anomalous otherwise as normal.

THE ANOMALY THRESHOLD (AT) AND MATURE ANTIGEN CONTEXT VALUE (MCAV)

Anomaly threshold (AT) is a default value that separates normal and abnormal antigens. It is used to compare the MCAV of an antigen. The antigen is abnormal/anomaly if the value exceeds the threshold value. Currently, there are three strategies to determine the AT for DCA; try and test experiment, class distribution between abnormal and normal group (Greensmith, 2007) and average MCAV (Song & Qijuan, 2012). The information in Table 1 summarizes the AT implementation in the existing work.

Table

Anomaly Detection Approach

Domain	Description	Dataset	Anomaly value	Approach		
				1	2	3
Network intrusion	Two steps. Firstly, the threshold is based on case of attack and total cases. Secondly, a new generated data (combining the attack mean and normal mean cases) will be added to old data and a new threshold will be generated to reduce error (Bukola & A.O., 2016)	NSHKDD	0.50%	×		
	The threshold value was based on the ratio between previous attack over overall data in order to determine denial of service attack (Gu, Greensmith, & Aickelin, 2008)	KDD99	0.80%		×	
	The MCAV represented the number of failure massages received in wireless network while the AT was determined through several experiments with expert guidance. The threshold value was within the value range 1-10 failure massage (Salmon et al., 2012)	Wireless network data generated by MICAz detector	1-10 failure massage	×		
	Three ATs were used based on the type of attack; Normality, Harmless Abnormality and Harm Abnormality. The values were determined based on several experiments. The value was determined after running several experiments (Chung-Ming & R., 2011)	Organization network data	0.50%	×		
Fault detection in robotic	The threshold value was the failure massage generated by a detector. The number of failure message is given by experts and then tested with several experiments within certain ranges.(Ran, et al., 2010)	Robot detector	800,600, -400, 200, 0,200,400,600,800, 1000,1500,2000	×		
E-mail classification	The threshold value was determined based on try and test which represented the number of spam e-mail (Secker, Freitas, & Timmis, 2003)	Spam e-mail record	0.50%	×		

(continued)

Domain	Description	Dataset	Anomaly value	Approach		
				1	2	3
Fraud detection	The threshold was based on the ratio of online fraud video rental over all transactions (Huang, Tawfik, & Nagar, 2010)	Online Rental video	0.28%		×	
Image classification	Modeling the type of leave and AT for the system was decided based on ratio of mature leave images over overall leave samples (Bendiab & Kholladi, 2011)	Leave images	-		×	
General classification	Introduce the mean MCAV as AT (Song & Qi-juan, 2012)	Breast cancer data	-			×

1- ‘try and test’, 2- class distribution, 3- min MCAV

The class distribution approach refers to the proportion between normal and abnormal classes where both classes need to be balanced in terms of number in order to produce a relational threshold value as depicted in Equation (3). This requirement is not easy to fulfill. Sometimes, since anomalies are isolated cases they tend to create a large gap between both classes.

$$AT_{\text{class distribution}} = (\sum \text{number of anomalies})/(\sum \text{total data points}) \qquad (3)$$

In the outbreak detection problem, for example, outbreak is a rare case that seldom occurs. It will cause the threshold value to be too small due to the big gap between the number of outbreak and non-outbreak cases. This can affect the detection accuracy as simulated in Table 2. Table 2 shows the result of DCA when AT is determined based on different class distribution ratios for breast cancer data (WBC). In the first row, the dataset was set to have a balance class between normal and abnormal patients while in the following row the number of abnormal patients was removed 90%. The result showed that the performance declined mainly at the ability to detect normal cases or lost its sensitiveness (SNS).

Table 2

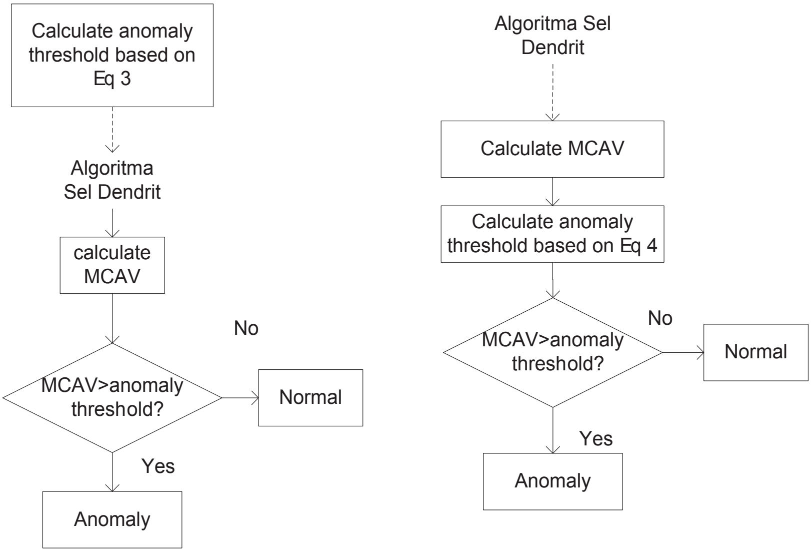
Anomaly Detection Problem Based on Class Distribution

	Total data		Threshold value	Result		
WBC	Normal	Anomaly		SNS	SPS	ACC
Original (100% anomaly)	241	458	0.65	0.976	1	0.984
Reduced (10% anomaly)	241	40	0.14	0.261	1	0.936

SNS- sensitivity, SPS- specificity, ACC-accuracy

The other issues with the existing implementations are that the value needs to be determined before mining based on historical information. The problem of this solution is the new data points tend to be unrecognizable if the pattern is distinct from the original setting. Besides that, the try and test approach is a time consuming process and highly depends on expert guidance. One of the solutions is calculating the value in real time during mining such min MCAV as shown in Equation (4) (Song & Qijuan, 2012). Although the mean MCAV approach able to skip the pre-determined AT, it has a drawback when facing extreme values among MCAV. Figure 1 shows the process of calculating the AT and comparing the value with MCAV using class distribution and min MCAV.

$$AT_{\text{min MCAV}} = (\sum \text{MCAV})/(\sum \text{total data points}) \tag{4}$$



(a) Class distribution (Greensmith, 2007) (b) Min MCAV (Song & Qijuan, 2012)

Figure 1. The steps of calculating AT and comparing it with MCAV based on class distribution and min MCAV.

THE PROPOSED METHOD

Two improvements were made in the proposed method, (a) normalizing the existing MCAV with upper CUSUM and (b) calculating new AT during real time by considering the acceptance value. Figure 2 shows the AT calculation step in DCA which was hybrid with CUSUM. The processes include calculating the average MCAV value, determining the acceptance value K, normalizing the MCAV with the upper CUSUM, and then comparing the normalized MCAV with the AT. This process started after DCA had calculated MCAV of its antigen. This improved algorithm is named NMZ_MCAV.

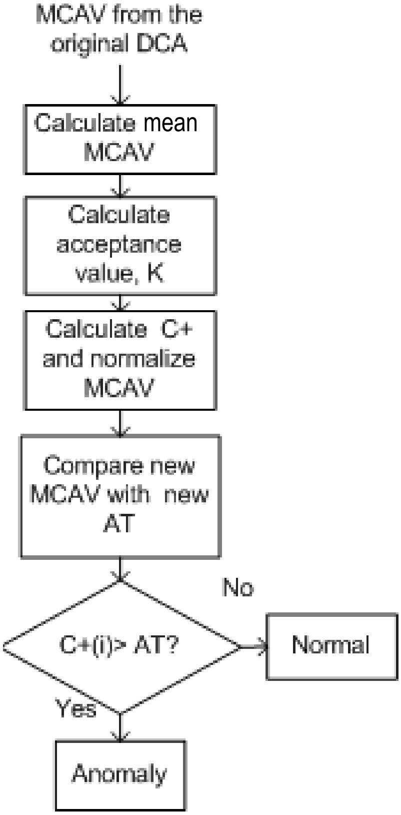

MCAV from the original DCA
Calculate mean MCAV
Calculate acceptance value, K
Calculate C+ and normalize MCAV
Compare new MCAV with new AT
C+(i)> AT?
No
Normal
Yes
Anomaly

Figure 2. The proposed NMZ_MCAV method.

Based on Figure 2, the input of this process is the MCAV which is generated from DCA learning. After calculating the mean MCAV, the acceptance value K is determined. K represents the allowable magnitude of change. It is expressed by Equation (5) where δ is the shift size from standard deviation σ. In this study, δ was set between 0-2 from the standard deviation σ.

$$K=\delta/2\ \sigma = (|\mu\_1-\mu\_0|)/2 \qquad (5)$$

Then, the upper side CUSUM is used to normalize MCAV. CUSUM is a statistical approach primarily used to monitor the planned process in manufacturing operations. It monitors the mean of the process and assumes a process remains under control when the cumulative mean is within the acceptance value K (Demsar, 2006). The process is considered out of control when a huge shift in movement occurs away from the target value. In this study, the cumulative mean shift was taken into consideration to normalize the MCAV. The upper side CUSUM, C+ was applied to normalize MCAV of each antigen such that

$$C_i^+ = max\ [0, x_i - \ (\mu_0 + K) + C_{i-1}^+] \tag{4}$$

where the $C_i^+$is the upper cumulative value at x$_{th}$ observation, x$_{\_i}$ is the process at i$_{th}$ observation, μ_0 is the initial mean and K is the allowance value which is chosen between the target and the out of control value μ_1. The C_i^+value accumulates deviation from μ_0 that is greater than K which is reset to zero on becoming negative. The starting value C_i^+=0.

After that is to obtain a new AT. In this step, the acceptance value is considered in the process by adding it with the existing mean MCAV such that

$$AT = mean\ MCAV + K \tag{5}$$

The function of K is to eliminate the existence of the extreme value in MCAV. Then, the final step is to compare the new MCAV and AT. Figure 3 depicts the proposed DCA enhancement algorithm.

```
Input: MCAV antigen, Magnitude of change, δ
Output: the normalized MCAV, new AT, final antigen status
START
  Calculate mean MCAV
  Normalize MCAV $c$
      Get μ and σ of all MCAV
      Calculate the acceptance value, K
      Normalize MCAV based on C+
   Calculate new anomaly threshold; $mean\ MCAV + K$
    Test the antigen abnormality status, if
         C+> AT = anomaly/abnormal
         C+< AT = normal
END
```

Figure 3. The proposed DCA Enhancement Algorithm

THE EXPERIMENT SETUP

This section discusses the experiment setup in order to evaluate the enhanced DCA algorithm. This proposed algorithm called NMZ_MCAV was compared with the existing DCA (M_MCAV) that is based on mean MCAV as AT strategy. Four evaluation metrics were applied, sensitivity (SNS), specificity (SPS), false detection rate (FDR), and accuracy (ACC). SNS measured the accurateness of the model to detect an abnormal class as an abnormal class; SPS measured the ability of the model to detect a normal class as a normal class; FDR measured the amount of false detections of an abnormal class as a normal class; and ACC measured the accurateness of the model in classifying both classes correctly. For SNS, SPS and ACC, the highest value indicated the best result while the lowest value was the best result for FDR.

In this study, 14 experiment datasets were used as described in Table 3. The first 12 datasets were benchmark or universal data from various domains that were downloaded from online data repositories. Meanwhile the last two datasets were outbreak datasets- dengue and respiratory-which were originally taken from the hospital and previous researchers. Both datasets were considered as case study in this study.

Table 3

Description of the Datasets

Dataset	Source	Data type	#Feature	#Record	#Target Class
Indian pima diabetic (DBC)	UCI (Murphy)	Transactional	9	768	2
Wisconsin breast cancer (WBC)			10	699	2
Iris (IRIS)			4	150	3
BUPA liver disorder (LDR)			7	345	2
Parkinson (PKN)			24	195	2
German credit (GCD)			25	1,000	2
Wine (WINE)			14	178	3
Biomedical (BIO)	StatLib archive (2005)		6	209	2

(continued)

Dataset	Source	Data type	#Feature	#Record	#Target Class
ECG	UCR Library (Award, 2008)	Time Series	100	101	2
Lightening (LTNG)			62	638	2
Coffee (CFE)			28	287	2
YOGA			301	427	2
Dengue	Vector Control Unit, Seremban, Malaysia		18	3,417	2
Respiratory	Wong et al. (2005)		12	23,645	2

Dengue dataset was provided by two departments; the emergency visit dataset from the Vector Control Unit, Seremban District Hospital, Negeri Sembilan, Malaysia and the climate dataset provided by the Meteorological Centre, Malaysia. The dataset was from 2003 to 2009. The emergency visit dataset had 15 features representing the demographic and clinical data of dengue patients. The climate dataset consisted of eight continuous attributes representing the information related to temperature, humidity and rain. Both datasets were then merged as one dengue profile dataset.

Respiratory was a synthetic dataset for influenza outbreak. Known as WSARE, this dataset was created by Wong (2004) for the outbreak detection model using the association rule and statistic. The dataset contained 100 sets of data with different outbreak patterns and the virus released date and WSARE7 was chosen for this study. The age of this dataset was from 2002 and 2003 with 12 categorical features and 23,647 daily data points.

RESULT AND FINDING

The performance of the proposed algorithm (NMZ_MCAV) is presented in this section. The enhanced algorithm NMZ_MCAV was compared with the existing DCA (M_MCAV) that used mean MCAV as AT. To present the result, this section is divided into two parts based on the benchmark dataset and the outbreak data.

Benchmark dataset

The benchmark dataset is a universal data of various domains that were downloaded from shared online data repositories. The evaluation results are shown in Table 4. In Table 4, each row represents the result of each dataset. The last the two rows summarize the average values of each performance metric and the results for all datasets in terms of wins, ties, and losses (indicated by W/T/L) towards 12 datasets. The W/T/L is considered in addition to the average measurement because the average criteria would be susceptible to outliers. The p value (pval) represents the significant test (Wilcoxon test or T-test), where the value of the NMZ_MCAV must be less than 0.05 to make it statistically significant compared to the M_MCAV (Demsar, 2006).

The results published in Table 4 indicate a positive improvement where NMZ_MCAV generates a superior result than M_MCAV in most datasets. The AVG score of each performance metrics show that the proposed approach has improved compared to competitor. The W/T/L statistics summarizes the capability of NMZ_MCAV to detect anomaly better that M_MCAV in most datasets. Although in certain datasets M_MCAV overcame NMZ_MCAV, their result was comparable and not significantly different.

Table 4

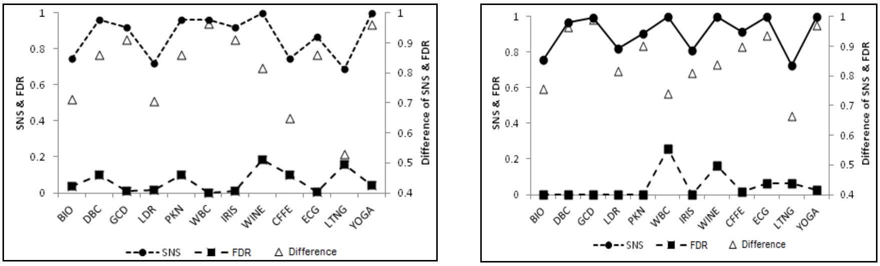
Comparative Results between NMZ_MCAV and M_MCAV for 12 Benchmark Datasets

	SNS				SPS			
	M_ MCAV	NMZ_ MCAV	△	pval	M_ MCAV	NMZ_ MCAV	△	pval
BIO	0.748	0.758	0.010^{W}	0.386^{W-}	0.964	0.999	0.035^{W}	0.000^{W+}
DBC	0.960	0.966	0.006^{W}	0.537^{T-}	0.900	1.000	0.099^{W}	0.000^{W+}
GCD	0.921	0.992	0.071^{W}	0.000^{W+}	0.991	0.999	0.008^{W}	0.000^{T+}
LDR	0.720	0.818	0.098^{W}	0.000^{W+}	0.986	0.998	0.012^{W}	0.000^{T+}
PKN	0.960	0.902	-0.058^{L}	0.000^{T+}	0.900	1.000	0.100^{W}	0.000^{W+}
WBC	0.964	1.000	0.036^{W}	0.000^{T+}	1.000	0.740	0.260^{L}	0.000^{T+}
IRIS	0.919	0.811	-0.109^{L}	0.000^{T+}	0.992	1.000	0.008^{W}	0.000^{T+}
WINE	1.000	1.000	0.000^{T}	-	0.815	0.838	0.023^{W}	0.000^{W+}
CFFE	0.749	0.916	0.167^{W}	0.000^{T+}	0.901	0.982	0.081^{W}	0.000^{T+}
ECG	0.867	1.000	0.133^{W}	0.000^{T+}	0.995	0.935	-0.061^{L}	0.000^{T+}
LTNG	0.688	0.726	0.038^{W}	0.037^{W+}	0.843	0.939	0.097^{W}	0.000^{T+}
YOGA	1.000	1.000	0.000^{T}	-$^{-}$	0.960	0.972	0.013^{W}	0.000^{T+}
AVG.	0.891	0.918			0.920	0.950		
W/T/L			8/2/2				10/0/2	

(continued)

	FDR				ACC			
	M_ MCAV	NMZ_ MCAV	△	pval	M_ MCAV	NMZ_ MCAV	△	pval
BIO	0.036	0.001	0.035^{W}	0.000^{W+}	0.886	0.913	0.026^{W}	0.000^{W+}
DBC	0.100	0.000	0.099^{W}	0.000^{W+}	0.921	0.988	0.067^{W}	0.000^{W+}
GCD	0.009	0.001	0.008^{W}	0.000^{T+}	0.970	0.997	0.027^{W}	0.000^{W+}
LDR	0.014	0.002	0.012^{W}	0.000^{T+}	0.832	0.894	0.062^{W}	0.000^{W+}
PKN	0.100	0.000	0.100^{W}	0.000^{W+}	0.921	0.926	0.005^{W}	0.013^{W+}
WBC	0.000	0.260	-0.260^{L}	0.000^{T+}	0.976	0.910	-0.066^{L}	0.000^{T+}
IRIS	0.008	0.000	0.008^{W}	0.000^{T+}	0.968	0.937	-0.031^{L}	0.000^{T+}
WINE	0.185	0.1615	0.023^{W}	0.000^{W+}	0.865	0.8821	0.017^{W}	0.000^{W+}
CFFE	0.099	0.018	0.081^{W}	0.000^{T+}	0.825	0.949	0.124^{W}	0.000^{T+}
ECG	0.005	0.065	-0.061^{L}	0.000^{T+}	0.949	0.958	0.009^{W}	0.000^{T+}
LTNG	0.158	0.061	0.097^{W}	0.000^{T+}	0.769	0.838	0.069^{W}	0.000^{W+}
YOGA	0.040	0.028	0.013^{W}	0.000^{T+}	0.964	0.975	0.011^{W}	0.000^{T+}
AVG.	0.080	0.050			0.897	0.935		
W/T/L			10/0/2				10/0/2	

Besides that, the NMZ_MCAV with the new AT has better ability to accurately detect anomaly as anomaly and at the same time can reduce error in misclassifying normal records as anomaly as this is an indicator of a good detection algorithm. Figure 4 summarizes the results in terms of SNS and FDR. The higher gap/range between both elements indicates the model is able to discriminate normal and abnormal groups effectively.

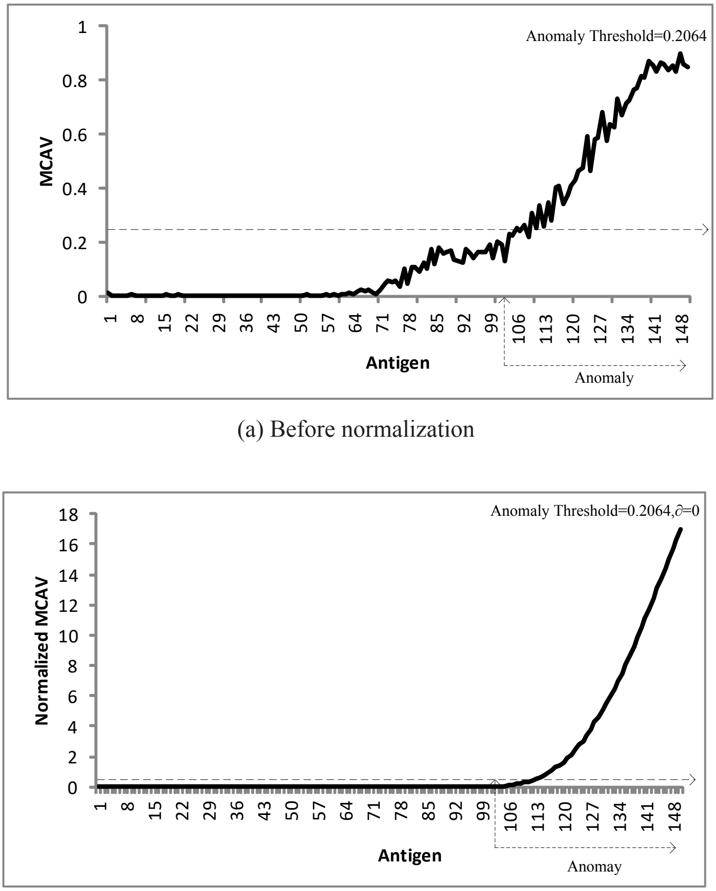


M_MCAV

NMZ_MCAV

Figure 4. The range between SNS and FDR for NMZ_MCAV and M_ MCAV in benchmark datasets.

Through the proposed approach, each antigen will have a new normalized MCAV and the value will be in a similar range with its neighbor if their characteristics are identical. Besides normalizing the MCAV with CUSUM, the acceptable value K in AT also can eliminate the existence of extreme MCAV values and this will improve detection accuracy. Figure 5 demonstrates the MCAV value before and after normalization using the proposed approach for IRIS dataset. It also shows that the MCAV of antigen before normalization does not consistently behave and the pattern changes into a uniform form after normalization.



(a) Before normalization

(b) After normalization

Figure 5. The MCAV of antigen before and after normalization with CUSUM in IRIS dataset.

Outbreak dataset

The performance of the proposed approach was then experimented with outbreak datasets- dengue outbreak and respiratory outbreak. Firstly, the enhanced algorithm NMZ_MCAV produced a better result than the previous model; M_MCAV in terms of SNS on both datasets as displayed in Figure 6. The accuracy of NMZ_MCAV increased by 0.07 and 0.18 for dengue and respiratory respectively when accurately classifying normal data as non-outbreak data that contributes to SPS score 0.814 (dengue) and 0.996 (respiratory).

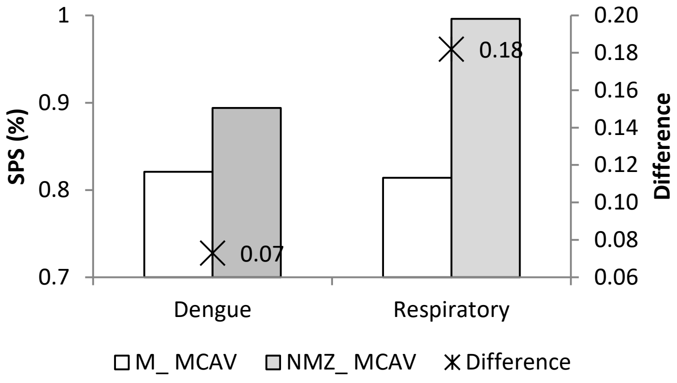


Figure 6. The SPS between NMZ_MCAV and M_MCAV on dengue and respiratory dataset.

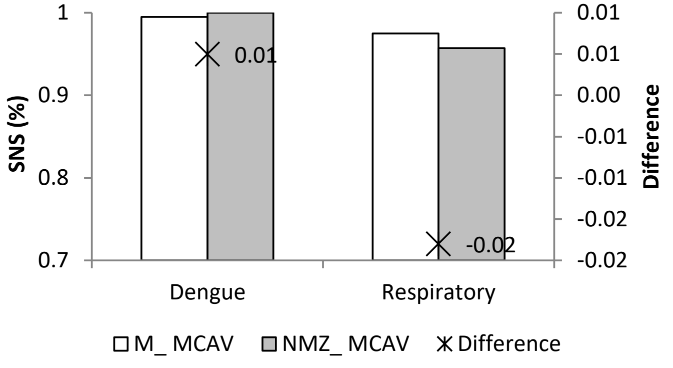


*Figure 7. The SNS b*etween NMZ_MCAV and M_MCAV on dengue and respiratory dataset.

In terms of the ability to detect the epidemic week or SNS, the NMZ_MCAV showed improvement (1.00) in comparison with M_MCAV (0.995). For respiratory data, the ability of NMZ_MCAV declined by 0.02 as compared to M_MCAV. However, their differences were small and not significant. Although the specificity result was slightly lower than its previous version, the proposed method had improved the ability of DCA in terms of sensitivity to detect the true outbreak week. Figure 7 shows a comparison between NMZ_MCAV and M_MCAV in terms of SNS on dengue and respiratory data.

The analysis was continued on the relationship between SNS and FPR over DCA after the MCAV was normalized with CUSUM. The comparison is shown in Table 5. Based on the table, NMZ_MCAV showed a better result in balancing the SNS (the ability to detect outbreak week as outbreak) and reducing the FPR (the error rates while detecting normal week as outbreak). The difference between both measurements shows the NMZ_MCAV performance was more consistent with higher SNS and lower FPR than M_MCAV. In addition, there were improvements in terms of average SNS and FPR for both sets.

Table 5

The Difference between SNS and FPR of NMZ_MCAV and M_MCAV for Dengue and Respiratory Dataset

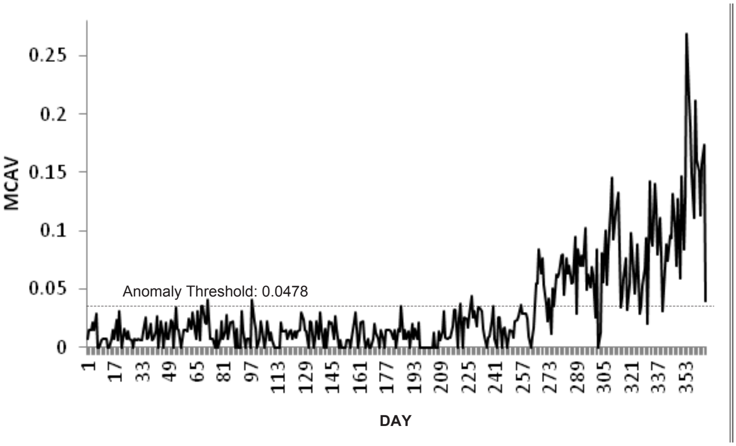
	M_ MCAV			**NMZ_ MCAV**		
	SNS	**FPR**	△	**SNS**	**FPR**	△
Dengue	0.995	0.179	0.816	1	0.106	0.894^{W}
Respiratory	0.975	0.186	0.789	0.957	0.004	0.953^{W}
Average	0.985	0.1825		0.9785	0.055	

Table 6

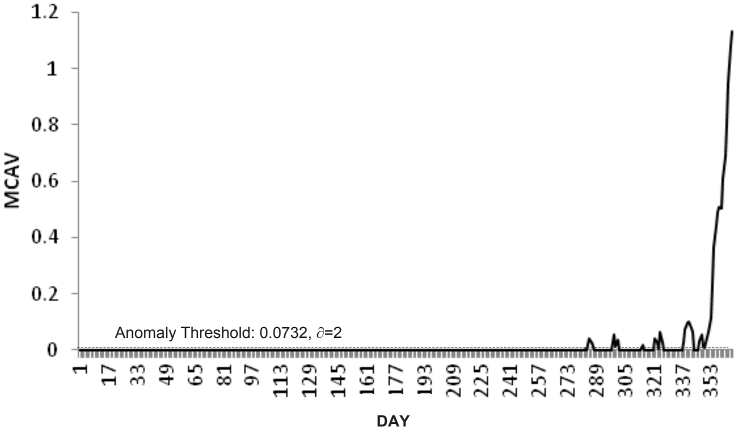
The ACC of NMZ_MCAV and M_MCAV for Dengue and Respiratory Dataset

	ACC		
	M_ MCAV	**NMZ_ MCAV**	△
Dengue	0.885	0.933	0.048 ^{W}
Respiratory	0.82	0.995	0.175 ^{W}
Average	0.897	0.935	

In addition, the ACC produced by NMZ_MCAV shows the proposed approach has helped DCA to increase the ACC for dengue data (0933) and respiratory (0995). On average, the ACC NMZ_MCAV was higher than the value produced by M_MCAV as shown in Table 6.



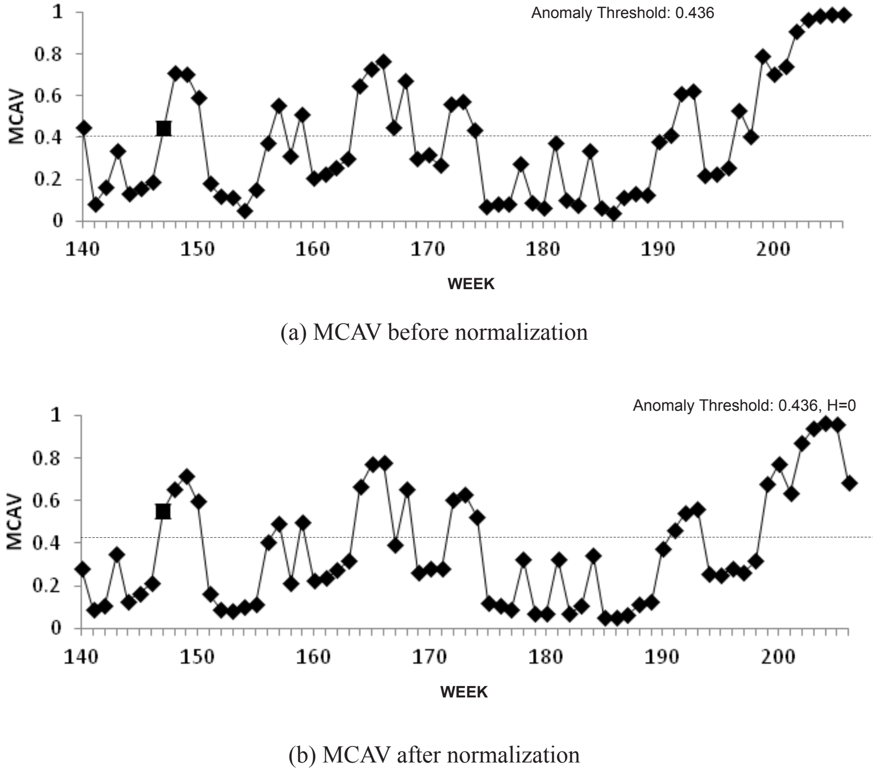
(a) MCAV before normalization



(b) MCAV after normalization

Figure 7. MCAV of respiratory dataset before and after normalization with CUSUM.

As in the benchmark data section, the proposed normalization using CUSUM will transform the MCAV from an inconstant pattern into a smaller and uniform value based on the similarity of the antigen characteristics. Figure 7 shows the MCAV before and after normalization for respiratory dataset and Figure 8 shows the dengue dataset. Based on both figures, the MCAV after normalization tends to have a uniform value than the previous model. For example, in respiratory dataset, the outbreak started on day 350 and remained for 14 days. The MCAV value before the outbreak remained low and suddenly spiked up on day 350. In comparison the MCAV value before normalization indicated an inconsistent pattern. For the dengue dataset as in Figure 8, the displayed MCAV value was for week 140 until week 203. In comparison with the respiratory dataset, it was noticed that the MCAV after normalization of the dengue dataset was not much different from before the normalization since its input signals were formalized according to the dengue definition given by the health ministry.



(a) MCAV before normalization

(b) MCAV after normalization

Figure 8. MCAV of dengue dataset before and after normalization with CUSUM.

From the experiments, it can be concluded that the performance of the DCA has increased in terms of SNS, SPS and ACC as well as the lower error rate when the MCAV has been normalized with CUSUM and consider the acceptance value K in the threshold value. Experiments on benchmark and outbreak datasets showed an improvement after its implementation. Table 7 below summarizes the differences between the DCA with normalized MCAV version (NMZ_MCAV) and without normalization (M_MCAV).

Table 7

A Comparison between NMZ_MCAV and M_MCAV

Feature	M_MCAV	NMZ_MCAV
MCAV characteristics	No relationship between neighboring antigens	There is a relationship between antigens. The neighboring antigen with similar characteristics will be normalized with similar MCAV value
MCAV value	Between 0 and 1	Depends on the accumulative min MCAV
Handling extreme value	No	Yes
Relevancy for unordered data	No	No
Classifier performance	Good	Better

CONCLUSION

An adaptive anomaly threshold for DCA called NMZ_MCAV was proposed in this paper. In the new approach, the upper CUSUM formula was used to normalize MCAV and then the new anomaly threshold was calculated during mining by considering the acceptance value K and min MCAV. By using the proposed solution, the performance of DCA was significantly improved in term of sensitivity, specificity, false detection rate, and accuracy after it was tested over 12 benchmark datasets and two outbreak datasets. In future, the NMZ_MCAV will be experimented on the real time network intrusion data and the business fraud data in order to further evaluate its effectiveness and robustness.

ACKNOWLEDGMENTS

The authors are grateful to the Seremban Public Health Department, Malaysia and the Faculty of Health Sciences, UKM, Malaysia as well as the Exploratory Research Grant Scheme for supporting this research project through grant ERGS/1/2011/STG/UKM/02/49.

REFERENCES

Anandita, S., Rosmansyah, Y., Dabarsyah, B., & Choi, J. U. (2015). *Implementation of dendritic cell algorithm as an anomaly detection method for port scanning attack*. Paper presented at the International Conference on Information Technology Systems and Innovation (ICITSI), Bandung.

Award, N. C. (2008). The UCR Time Series Classification/Clustering Page Retrieved 6 Januari 2014, 2014, from http://www.cs.ucr.edu/~eamonn/time_series_data/#SwedishLeaf

Bendiab, E., & Kholladi, M. K. (2011). Recognition of plant leaves using the dendritic cell algorithm. *International Journal of Digital Information and Wireless Communications (IJDIWC), 1*(1), 284-292.

Bukola, O., & A.O., A. (2016). Auto-immunity dendritic cell algorithm. *International Journal of Computer Applications, 2*(137), 10-17.

Chelly, Z., & Elouedi, Z. (2016). A survey of the dendritic cell algorithm. *Knowl. Inf. Syst., 48*(3), 505-535. doi: 10.1007/s10115-015-0891-y

Chung-Ming, & R., O. C. (2011, Aug. 29 2011-Sept. 1 2011). Immunity-inspired host-based intrusion detection systems. *Proceeding of the The Fifth International Conference on Genetic and Evolutionary Computing (ICGEC)*, Kinmen, Taiwan.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res., 7*(1), 1-30.

El-Alfy, E.-S. M., & AlHasan, A. A. (2016). Spam filtering framework for multimodal mobile communication based on dendritic cell algorithm. *Future Generation Computer Systems, 64*, 98-107. doi: http://dx.doi.org/10.1016/j.future.2016.02.018

Greensmith, J. (2007). The dendritic cell algorithm (Unpublished doctoral dissertation), University of Nottingham, United Kingdom.

Greensmith, J., Aickelin, U., & Cayzer, S. (2005). Introducing dendritic cells as a novel immune inspired algorithm for anomaly detection. *Proceeding of the 4th International Conference in Artificial Immune Systems (ICARIS)*, Banff, Alberta, Canada.

Greensmith, J., Twycross, J., & Aickelin, U. (2006). Dendritic cells for anomaly detection. *Proceeding of the The IEEE Congress on Evolutionary Computation (CEC) Vancouver*, BC, Canada.

Gu, F., Greensmith, J., & Aickelin, U. (2008). Further exploration of the dendritic cell algorithm: Antigen multiplier and time windows. In P. Bentley, D. Lee & S. Jung (Eds.), *Artificial Immune Systems* (Vol. 5132, pp. 142-153): Springer Berlin Heidelberg.

Huang, R., Tawfik, H., & Nagar, A. K. (2010, 23-26 Sept. 2010). On the use of innate and adaptive parts of artificial immune systems for online fraud detection. *Proceeding of the The Fifth IEEE International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA)* Changsha, China.

Lee, N. M. Y., Lau, H. Y. K., Wong, R. H. K., Tam, W. W. L., & Chan, L. K. Y. (2016). Dendritic cells for behaviour detection in immersive virtual reality training. *Proceeding of the International Conference on Innovative Techniques and Applications of Artificial Intelligence (SGAI 2016)*, Chambrige, England.

Mohamad Mohsin, M. F., Hamdan, A. R., & Abu Bakar, A. (2013). The preliminary design of outbreak detection model based on inspired immune system. *Proceeding of the Third World Congress on Information and Communication Technologies (WICT 2013)*, Hanoi, Vietnam.

Murphy, P. M. (1997). UCI repositories of machine learning and domain theories" Retrieved 2 January 2013, from http://www.ics.uci.edu/~mlearn/MLRepository.html

Ou, C.-M. (2012). Host-based intrusion detection systems adapted from agent-based artificial immune systems. *Neurocomputing, 88*(0), 78-86.

Ran, B., Timmis, J., & Tyrrell, A. (2010, 18-23 July 2010). The diagnostic dendritic cell algorithm for robotic systems. *Proceeding of the IEEE Congress on Evolutionary Computation (CEC)*, Barcelona, Spain.

Salmon, H., Farias, C., Loureiro, P., Pirmez, L., Rossetto, S., Rodrigues, A., . . . Costa Carmo, L. (2012). Intrusion detection system for wireless sensor networks using danger theory immune-inspired techniques. *International Journal of Wireless Information Networks*, 1-28.

Secker, A., Freitas, A., & Timmis, J. (2003). A danger theory inspired approach to web mining. In J. Timmis, P. Bentley & E. Hart (Eds.), *Artificial Immune Systems* (Vol. 2787, pp. 156-167): Springer Berlin Heidelberg.

Song, Y., & Qi-juan, C. (2012, 25-27 May 2012). A dendritic cell algorithm for real-time anomaly detection. *Proceeding of the IEEE International Conference on Computer Science and Automation Engineering (CSAE)*, Zhangjiajie, China.

Song, Y., & Qijuan, C. (2012, 26-27 Aug. 2012). Dendritic cell algorithm for anomaly detection in unordered data set. *Proceeding of the 4th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Nanchang, China.

StatLib. (Jul 19, 2005). *Statlib — datasets archive*. Retrieved 3 Febuary 2014, 2013, from http://lib.stat.cmu/datasets

Wong, W.-K. (2004). *Data mining for disease outbreak detection* (Unpublished doctoral dissertation). Carnigie Mallon University, Pittsburgh.

Wong, W.-K., Moore, A., Cooper, G., & Wagner, M. (2005). What's strange about recent events (WSARE): An algorithm for the early detection of disease outbreaks. *Journal of Machine Learning Research, 6*, 1961--1998.

Zainal, K., & Jali, M. Z. (2017). The significant effect of feature selection methods in spam risk assessment using dendritic cell algorithm. *Proceeding of the Fifth International Conference on Information and Communication Technology (ICoICT)*, Malacca, Malaysia.