How to cite this paper:

Mohamad Farhan Mohamad Mohsin, Azuraliza Abu Bakar, & Abdul Razak Hamdan. (2017). An adaptive anomaly threshold in artificial dendrite cell algorithm in Zulikha, J. & N. H. Zakaria (Eds.), Proceedings of the 6th International Conference of Computing & Informatics (pp 250-255). Sintok: School of Computing.

AN ADAPTIVE ANOMALY THRESHOLD IN ARTIFICIAL DENDRITE CELL ALGORITHM

Mohamad Farhan Mohamad Mohsin[1], Azuraliza Abu Bakar[2], and Abdul Razak Hamdan[2]

[1]*School of Computing, Universiti Utara Malaysia, farhan@uum.edu.my*
[2] *Faculty of Science & Information Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia*, aab@ukm.my, *arh@ftsm.ukm.my*

ABSTRACT. The dendrite cell algorithm (DCA) relies on the multi-context antigen value (MCAV) to determine the abnormality of a record by comparing it with anomaly threshold. In practice, the threshold is pre-determined before mining based on previous information and the existing MCAV is inefficient when expose to extreme values. This causes the DCA fails to detect unlabeled data if the new pattern distinct from previous information and reduces the detection accuracy. This paper proposed an adaptive anomaly threshold 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 datasets, the new version of DCA generated a better detection result than its previous version in term of sensitivity, specificity, false detection rate, and accuracy.

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

INTRODUCTION

The dendritic cell algorithm (DCA) is a biologically-inspired algorithm that belongs to the artificial immune system (AIS). It replicates the nature behavior of human defense mechanism against pathogen based on the concept of danger theory, which believes the human immune system is triggered when a dendritic cell recognizes a danger signal released by an unexpected cell death due to pathogenic infection. In the same way as it is the responsibility of the dendritic cell to recognize an intruder (bacteria, virus, and parasite) that enters the body, the DCA is modeled to detect anomalies mainly in time series related applications. Since it was introduced in 2005 (Greensmith, Aickelin, & Cayzer, 2005), DCA has been widely applied in various areas, mainly to time series anomaly detection-based problems including fault (Ran, Timmis, & Tyrrell, 2010), fraud (Huang, Taufik, & Nagar, 2009), intrusion (Ou, 2012), and outbreak (Mohamad Mohsin, Abu Bakar, & Hamdan, 2014) detection. The published results of these applications demonstrate that DCA performs well in terms of detecting hidden anomalies in comparison to other systems.

The distinct advantages of DCA over other data mining approaches is the anomalies detection mechanism where it employs the dangerousness of an antigen, known as the multi-context antigen value (MCAV), rather than a pattern-matching approach. During monitoring, each record is considered as antigen that has possibility to be attacked by pathogen and the health information of each antigen during its life span is recorded as MCAV. MCAV is the

final contact that represents antigen experience throughout its life span based on the frequency of mature antigen over total antigen. At the end, the antigen is labeled as abnormal if the MCAV is greater than the pre-determined anomaly threshold.

In practice, there were several approaches in determining the anomaly threshold; the try and test which is based on expert recommendation, the class distribution between abnormal and normal group (Greensmith, 2007), and lastly the mean MCAV(Song & Qijuan, 2012). Their limitation is the threshold values is need to be determined before mining based on historical information that can cause the unseen new record tend to be unrecognizable if the pattern distinct from the original setting. To avoid this, it is an advantage if the threshold can be calculated in a real time during mining. Although in the mean MCAV approaches, it able to skip the pre-determine anomaly threshold, it has drawback when facing with extreme values among MCAV. Therefore, an adaptive anomaly threshold in DCA based on Cumulative Sum (CUSUM) approach was proposed in this paper. The aim of the enhancement was to improve DCA that can determine threshold value during mining stage, robust against extreme value so that it can increase its performance in detecting anomaly. The proposed algorithm was compared with the previous DCA with mean MCAV and four evaluations criteria were applied; the sensitivity, specificity, false detection rate, and accuracy. In this study, 12 universal datasets from several data provider were chosen as experiment data.

The remainder of this paper is organized as follows: It starts with highlighting the dendrite cell algorithm background. Next is the presentation of results and discussion, and finally is 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 is activated when a body cell releases danger signal as response to infection (Matzinger, 2012). Biologically, the main element of the theory, the DCs will recognizes the released signals by collecting body cells protein paired with three signals; PAMP, DS, SS and then monitors their life progress. The monitoring task continues until the cell dies either as a 'healthy death' (normal) or 'unhealthy death' (abnormal).

Analogized from 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 dataset is marked as antigen that has chances to be attack by pathogen. 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 multiple antigens sampling, calculates the changes and determines which antigen is causing the changes using the accumulative function such that $O_j(x) = (\sum_{i=0}^{i=3} W_{ij} * IS_{ij}(x))) / (\sum_{1=0}^{i=3} |W_{ij}|)$ 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 categories CSM, Mature, and Semi-Mature.

All input signals are transformed into three cumulative output signals: CSMs, Mature, and Semi-Mature. Throughout several sampling, the output signals will change the immature DCs 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 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 end. 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 MCAV = (Mature)/(Semi Mature + Mature). If the MCAV is above a predetermined value (anomaly threshold), the antigen is label as abnormal/anomalous otherwise as normal.

THE PROPOSED METHOD

Anomaly threshold (AT) is a value that separates normal and abnormal antigen. It is used to compare the MCAV of each antigen such that the antigen is abnormal/anomaly if the value is bigger than threshold. There were two improvements made in the proposed method; (1) normalizing the existing MCAV with upper CUSUM and (2) calculating new AT real time by considering the acceptance value. Figure 1 shows the AT calculation step in DCA which was hybrid with CUSUM. The processes include calculating the average MCAV value, determine the acceptance value K, normalize the MCAV with the upper CUSUM, and then comparing the normalized MCAV with the AT. This process started after DCA had calculated MCAV of ts antigen. This improve algorithm is named as NMZ_MCAV.

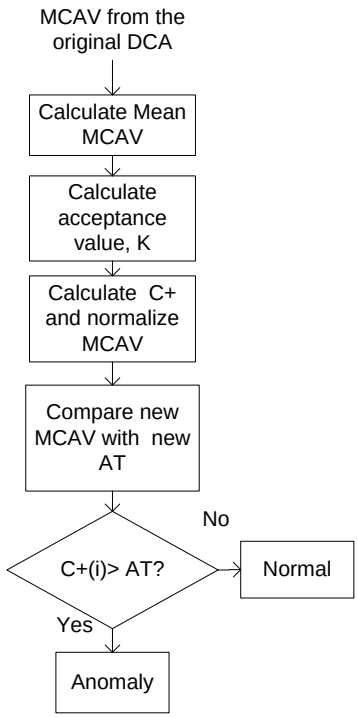


Figure 1. The proposed adaptive anomaly threshold based on CUSUM (NMZ_MCAV)

Based on Figure 1, 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 K=δ/2 σ = ($|\mu_1$-$\mu_0|$)/2 , where δ is the shift size from standard deviation, σ. In this study, δ was set between 0-2 from standard deviation σ.

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 control value. 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 is taken into consideration to normalize the MCAV. The upper side CUSUM, $C+$ is applied to normalize MCAV of each antigen such that $C_i^+ = max\,[0, x_i - (\mu_0 + K) + C_{i-1}^+]$ where the $C_i^+$is the upper cumulative value at i_{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 $\mu_0$and 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 new AT. In this step, the acceptance value is considered in the process by adding it with existing mean MCAV such that $AT = mean\ MCAV + K$. The function of K is to eliminate the existence of extreme value in MCAV. Then, the final steps is to compare the new MCAV and AT. Figure 2 depicts the proposed DCA enhancement algorithm.

Input: MCAV antigen, Magnitude of change, δ

```
Output: the normalize MCAV, new AT, final antigen status
0  START
1    Calculate mean MCAV
2    Normalize MCAV c
3        Get μ and σ of all MCAV
4        Calculate the acceptance value, K
5        Normalize MCAV based on C+
6     Calculate new anomaly threshold; mean MCAV + K
8    Test the antigen abnormality status, if
9            C+> AT = anomaly/abnormal
10           C+< AT = normal
11 END
```

Figure 2. The proposed DCA enhancement algorithm

RESULT AND FINDING

In this section, the performance of the proposed algorithm (NMZ_MCAV) is presented. This enhanced algorithm NMZ_MCAV was compared with the existing DCA (M_MCAV) that used mean MCAV as AT. 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. There were 12 experiment datasets used in this study as described in Table 1.

Table 1. Description of the datasets

Dataset	Origin	Attributes #	Records #
Indian Pima Diabetic (DBC)	Murphy (1997)	9	768
Wisconsin Breast Cancer (WBC)		10	699
Iris (IRIS)		4	150
BUPA Liver Disorder (LDR)		7	345
Parkinson (PKN)		24	195
German Credit (GCD)		25	1000
Wine (WINE)		14	178
Biomedical (BIO)	StatLib (2005)	6	209
Coffee (CFFE)	Award (2008)	28	287
ECG (ECG)		100	101
Lightening (LTNG)		62	638
Yoga (YOGA)		301	427

The evaluation results were demonstrated in Table 2. In Table 2, each row represents the result for each dataset. The last the two rows summarize (1) the average values of each performance metric and (2) the results for all datasets in term 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 2 indicate a positive improvement where NMZ_MCAV generates superior result than M_MCAV in most datasets. The AVG score of each performance metrics show that the proposed approach has improved from its competitor. In the W/T/L statistics, it summarizes the capability of NMZ_MCAV to detect anomaly better that

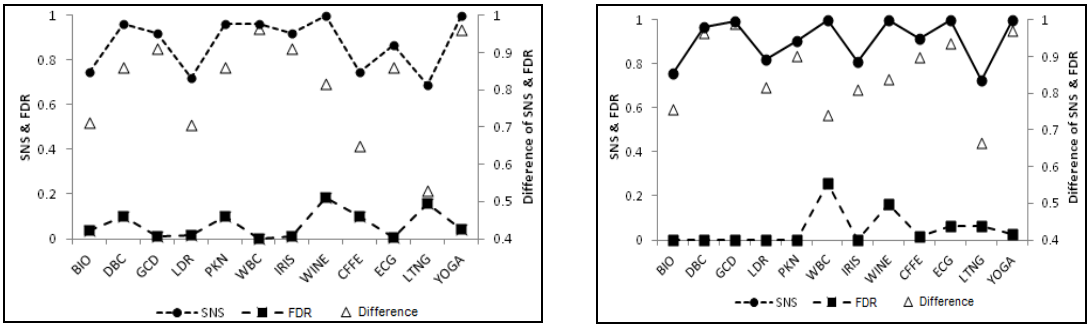
M_MCAV in most datasets. Although in certain datasets M_MCAV overcome NMZ_MCAV, their result is comparable and not significantly difference.

Table 2. Comparative results of NMZ_MCAV and M_MCAV for 12 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	

	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 new AT is proven has better ability when able to accurately detect anomaly as anomaly and in the same time can reduce error in misclassifying a normal records as anomaly as this is a indicator of a good detection algorithm. Figure 3 summarizes the results in term of SNS and FDR such that the higher gap/range between both elements indicates the model able to discriminate normal and abnormal group effectively.



M_MCAV NMZ_MCAV

Figure 3. The range between SNS and FDR for NMZ_MCAV and M_MCAV

CONCLUSION

This paper proposed an adaptive anomaly threshold for DCA using CUSUM approach. In the proposed approach NMZ_MCAV, the existing MCAV were normalized with upper CUSUM and the new AT was calculated during mining by considering the acceptance value K and min MCAV. From the experiments over 12 datasets, the new DCA version generated a better detection result than M_MCAV in term of sensitivity, specificity, false detection rate, and accuracy. To further evaluate the effectiveness of the proposed approach, further analysis will be conducted on real world data such disease outbreak and computer monitoring.

REFERENCES

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

Demsar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.*, *7*(1), 1-30.

Greensmith, J. (2007). *The Dendritic Cell Algorithm.* Ph.D, University of Nottingham, UK.

Greensmith, J., Aickelin, U., & Cayzer, S. (2005). *Introducing Dendritic Cells as a Novel Immune Inspired Algorithm for Anomaly Detection* Paper presented at the 4th International Conference in Artificial Immune Systems (ICARIS).

Huang, R., Taufik, H., & Nagar, A. K. (2009). *Artificial Dendritic Cells Algorithm for Online Break-in Fraud Detection*. Paper presented at the Second International Conference on Development in eSystems Engineering.

Matzinger, P. (2012). The evolution of the danger theory. *Expert Review of Clinical Immunology*, *8*(4), 311-317.

Mohamad Mohsin, M. F., Abu Bakar, A., & Hamdan, A. R. (2014). Outbreak detection model based on danger theory. *Applied Soft Computing*, *24*(0), 612-622.

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). *The Diagnostic Dendritic Cell Algorithm for robotic systems.* Paper presented at the IEEE Congress on Evolutionary Computation.

Song, Y., & Qijuan, C. (2012, 26-27 Aug. 2012). *Dendritic Cell Algorithm for Anomaly Detection in Unordered Data Set.* Paper presented at the 4th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC).

StatLib. (Jul 19, 2005). "Statlib- datasets archive" Retrieved 3 Feb14 from http://lib.stat.cmu/datasets