An Evaluation of Feature Selection Technique for Dendrite Cell Algorithm

Mohamad Farhan Mohamad Mohsin
School of Computing, College of Arts & Sciences
Universiti Utara Malaysia
Kedah, Malaysia
farhan@uum.edu.my

Abdul Razak Hamdan[1], Azuraliza Abu Bakar[2]
Data Mining and Optimization Research Group
Centre for Artificial Intelligence Technology, Faculty of Science & Information Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia.
arh@ftsm.ukm.my[1], aab@ftsm.ukm[2]

***Abstract*—Dendrite cell algorithm needs appropriates feature to represents its specific input signals. Although there are many feature selection algorithms have been used in identifying appropriate features for dendrite cell signals, there are algorithms that never been investigated and limited work to compare performance among them. In this study, six feature selection algorithms namely Information Gain, Gain Ratio, Symmetrical Uncertainties, Chi Square, Support Vector Machine, and Rough Set with Genetic Algorithm Reduct are examined and their effectiveness to represent dendrite cell signal are evaluated. Eight universal datasets are chosen and assessing their performance according to sensitivity, specificity, and accuracy. From the experiment, the Rough Set Genetic Algorithm reduct is found to be the most effect feature selection for dendrite cell algorithm when it generates a consistent result for all evaluation metrics. In single evaluation metrics, the chi square technique has the best competence in term of sensitiveness while the rough set genetic algorithm reduct is good at specificity and accuracy. In the next step, further analysis will be conducted on complex dataset such as time series data set.**

Keywords—artificial immune system; danger theory; dendrite cell algorithm; feature selection; signal selection

I. INTRODUCTION

The dendrite cell algorithm (DCA) is a computational system designed on the principles of natural immune system termed the danger theory. Under the theory, the dendrite cell (DCs) plays the most prominent roles in capturing infectious body cells and the biological DCs function is replicated into DCA as an agent to detect anomalies [1]. Within computer context, DCA transforms a set of data items into appropriate input signals asynchronously with location markers in the form of antigen to perform antigen classification [2]. At the end of signal processing, DCA determines the antigens abnormality based on the dangerousness of an antigen, known as the multi-context antigen value (MCAV). This characteristic makes DCA differ from existing anomaly detection approaches that usually apply pattern matching to identify an anomaly.

For functioning the algorithm, DCA relies on three specific signal types namely pathogen-associated molecular patterns (PAMP), danger signals (DS), and safe signals (SS). According to biological studies, each signal carries different characteristics and give dissimilar effects towards cell behavior that cause the healthiness of body cell vary. Firstly is PAMP signal that carries information about anomalous situation while the presence of SS indicates no anomalous situation. Meanwhile, the DS characteristic is it may or may not indicate an anomalous situation but the probability of an anomaly is higher than in the normal situation. In DCA, those signals are represented by set of appropriate features/attributes that carries similar characteristics as DS, SS, and PAMP signal. Therefore, it is an essential task to correctly map features and DCA signals otherwise high detection errors are produced.

Signal selection and signal normalization are two vital data pre-processing activates in DCA. In signal selection, it involves two tasks which are extracting suitable features and assigning them into appropriate DCA signals. In this phase, the feature selection technique is applied to select the most appropriate feature to represent the signals. After that, the assigned feature and signal is mapped and then normalized using specific normalization algorithm as it has relation to DCA signal characteristic.

In recent years, there is several future selection techniques have been applied in DCA including statistical analysis, principal component based analysis (PCA) and the information gain based method. Under the exploratory statistical analysis, Greensmith [2] used the standard deviation to filter the important feature for classifying breast cancer dataset. The features with higher standard deviation are chosen to represent PAMP, SS, and DS. In [3], the information gain approach is applied to select DCA signals for KDD 99 dataset. Based on information gain value, 10 network features are selected to be DCA signals where 5 features is set as PAMP, 3 features as SS, and 2 features as DS. Then, the value of this attribute is then normalized into the range from 0 to 100 using linear normalization appraoch. In different approach, Gu et. al [4] proposed the PCA to reduce features in biometrics dataset for the stress recognition of automobile drivers and then map to DCA signal. Their result is successful but PCA seems has several shortcomings towards other DCA application. For example PCA only well functions for a linear problem that cause unreliable results for non-linear problem. Besides that, the original dataset should be normalized in between 0 and 1

that might possibly affect the final result. Meanwhile, [5] demonstrate a new feature selection method for DCA using rough set approach called RC-DCA. In their work, the full reduct algorithm of rough set based is used to filter the important feature for DCA and gain better result than PCA. Later, they improved the RC-DCA model by using Quick Reduct algorithm (QR-DCA) with a higher accuracy result [6].

Although there are number of technique have been proposed to select appropriate feature for DCA, there are other approaches that never been investigated as well as limited work to compare among them. In this study, we aim to examine the performance of different feature selection algorithms towards DCA and evaluate their effectiveness in representing the signals. We choose six feature selection algorithms to be experimented namely Information Gain (IG), Gain Ratio (GR), Symmetrical Uncertainties (SU), Chi Square (CS), Support Vector Machine (SVM) and Rough Set with Genetic Algorithm Reduct (GA). To evaluate the algorithms, eight universal classification datasets are chosen and assessing their performance according to three measurement metrics: sensitivity, specificity, and accuracy.

This paper is organized as follows. Section II introduces the overview of DCA. Then, the feature selection technique used in this study is discussed in Section III. It will be followed by experiment setup in Section IV. In Section V, the finding of the study will be presented. The final sections conclude this work.

II. DENDRITE CELL ALGORITHM OVERVIEW

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. 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 in Equation 1.

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

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 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.

III. FEATURE SELECTION TECHNIQUE

Feature selection is frequently used as a data preprocessing step to data mining and machine learning. Its principal goal is to improve the mining accuracy by choosing only a subset of relevant and ignore non relevant features without decreasing the final result. Feature selection has three approaches to seek the most appropriate approaches; wrapper, filter, and hybrid between them. The output presentation has two; 1) a set of reduced feature 2) a feature ranking according its importance. The design and the formula of each feature selection method used in this study are briefly described.

A. Information Gain(IG)

The IG is an entropy-based feature evaluation method commonly used in machine learning and information theory. The goal is to measure the number of bits of information about the class prediction by knowing the presence or absence of a feature and the corresponding class distribution. A score of each feature is calculated depending on how much more information is gained with respect to the class. The IG formula is given as follows: IG(X) = H(Y) - H(Y|H) where H(Y) and H(Y|H) are the entropy of Y and the conditional entropy of Y given X, respectively. The merit of the feature is determined by how far is the reduction in entropy of the class when considered with the corresponding feature individually.

B. Gain Ratio (GR)

The GR is an improvement of IG method. In comparison to IG that is biased towards features with high values, the GR is aimed to maximize the feature's information gain and minimize the value of its value simultaneously. The GR formula is given as: GR(X) = ID(X) / IV(X) and IV(X) = $-\sum_{i=0}^{R}$(Xi|/N) Log (Xi|/N) where $|X_i|$ is the number of instances where feature X takes the value of X_i, r is the number of unique values of X, and N is the sum of items in data.

C. Rough Set Genatic Algorithm Reduct (GA)

The goal of reduct computation in rough set classifier is to select the most important that can be used to represents the decision system. Given $A = (U, A)$, a feature a is said to be dispensable in $B \subseteq A$ if indiscernibility relation $IND(B) = IND(B - \{a\})$ otherwise the feature is indispensable in B. Given an information system IS $A = (U, A)$ let $B \subseteq A$. A reduct of B is a set of feature $B' \subseteq B$ such that all features $a \in b - B'$ are dispensable and $IND(B) = IND(B')$. For GA reduct, it is the evolutionary algorithm that computes both

single and all reducts item for decision system with several different fitness functions and data representation.

D. Chi-square (CS)

The CS is a non-parametric statistical method used to determine if a distribution of observed frequencies differs from the theoretical expected frequencies. The goal of CS is to summarizes the discrepancies between the expected number of times each outcomes occurs and the observed number of times each outcome occurs, by summing the squares of the discrepancies, normalized by the expected numbers, over all the categories (REF). The chi-square formula is given as: $X^2 = \text{Sigma}[(O-E)^2/E]$ such that X^2 is the chi-square statistic, O is the observed frequency and E is the expected frequency.

E. Symmetrical Uncertainty (SU)

Symmetry is a desired property for a measure of correlations between features. The idea of SU is the improve drawback in IG which is biased in favor of features with more values. It evaluates the worth of an attribute by measuring the symmetrical uncertainty with respect to the class. The SU is defined as [7] $SU(X,Y)=2 * [IG(X|Y) / (H(X) + H(Y))]$ where $H(X)=-\sum_i (P(x_a) \text{Log}_2 (P(x_i)))$ is the entropy of feature X and $IG(X \mid Y) = H(X) - H(X|Y)$ is the information gain from X provided by class Y.

F. Support Vector Machine (SVM)

SVM is a wrapper based typed method that evaluates the worth of a feature by using SVM classifier by minimizing generalization bounds via gradient descent. It attempts to find the function from the set $f(x, w, b, \sigma) = w.\Phi(x.\sigma) + b$ that minimizes generalization error. The kernel function is used to maps item into a high dimensional space and construct an optimal hyperplane in the space.

IV. EXPERIMENT SETUP

Six feature selection algorithms are chosen to be presented into DCA and from that, the most effective feature selection technique for DCA is identified. To experiment them, every algorithms are evaluated by applying them to eight universal classification datasets, seven of which are taken from UCI Machine Learning Respiratory [8] and one from the StatLib Archive [9], as described in Table I.

TABLE I. EXPERIMENT DATASET

Dataset	Feature #	Records #	Decision Class #	Origin
Indian Pima Diabetic (DBC)	9	768	2	
Wins. Breast Cancer (WBC)	10	699	2	
Iris (IRIS)	4	150	3	
BUPA Liver Disorder (LDR)	7	345	2	[8]
Parkinson (PKN)	24	195	2	
German Credit (GCD)	25	1000	2	
Wine (WINE)	14	178	3	
Biomedical (BIO)	6	209	2	[9]

Firstly, the data is pre-processed, where the numerical missing value is replaced with a mean value while Mod for categorical feature. Since all dataset is unordered dataset, they need to be sorted in ascending order format according to decision class. This is to suit DCA's requirement as time series application where it only can be applied on timely ordered datasets. Besides that, the datasets is set to have two decision classes: abnormal and normal. To assign the selected features to appropriate DCA signals, we use the feature ranking generated by feature selection algorithm. Based on the highest ranking, only four attributes are chosen for DCA where attribute at rank 1 and rank 2 is set as DS and PAMP while the rest attributes are set as DS. After that, the attributes are normalized according DCA signal. The cumulative sum normalization technique is adopted to normalized DCA signal [10] .

For DCA, the initial parameter setting is formalized as follows. In all experiments, a population of 100 cells is created and the total cycle cell update is set to 20. In every cycle, DCs are allowed to perform antigen sampling10 times. The weight for the accumulative function is set to W1 = 1 and W2 = 2. The experiment is repeated 100 times and the average of each evaluation metric is recorded for analysis.

To evaluate the DCA performance, we examine the algorithms' results using three evaluation metrics: Sensitivity (SNS), Specificity (SPS), and Accuracy (ACC). SNS measure the accurateness of the model to detect abnormal class as abnormal class (SNS= TP/(TP+FN)) and the ability of the model to detect normal class as normal class is measured by SPS (SPS= TN/(TN+FP)). The ACC checks the accurateness of the model in classifying both classes correctly (ACC= (TP+TN)/(TP+TN+FN+FP)). The highest value of SNS, SPS, and ACC indicates the best result.

V. FINDING

This section presents the performance of the DCA when different feature selection algorithms are presented to it. The evaluation results in term of sensitivity (SNS), specificity (SPS), and accuracy (ACC) are demonstrated in Table II-Table IV. Each row represents the result for each dataset (BIO, LDR, DBC, GCD, IRIS, WBC, PKN, and WINE) while each column represents one of the six feature selection algorithms assessed in the study (GR, IG, CS SVM, SU, and GA). The last two rows is the average (AVG) that summarizes the total score of every feature selection algorithms towards eight datasets. Meanwhile the last row represents the frequency of algorithms (♘) become the highest model. From these results, it can be seen that each feature selection algorithms generates a good detection results in each evaluation metric.

The first analysis is the sensitiveness of DCA that measure its accurateness to detect anomalous item correctly as anomalous. As shown in Table II, each feature selection algorithm demonstrates a comparable result with no significant different for most datasets mainly the BIO, IRIS, GCD, WBC, and WBC. In certain datasets such GCD, the GR algorithm however indicates the worst result in comparison to other algorithms as well as the SU and GA. Meanwhile in LDR dataset, GA generates higher significant SNS than other algorithms. In overall, the CS algorithm seems to have better ability in recognizing abnormal item based on its highest

averages score (0.923) and lead the highest rank as the best model in most datasets.

TABLE II. DCA SENSITIVENESS TOWARDS FEATURE SELECTION

	GR	IG	CS	SVM	SU	GA
BIO	0.738	0.803✓	0.797	0.754	0.799	0.736
IRIS	0.947✓	0.941	0.944	0.945	0.945	0.919
WINE	0.999	0.996	0.998	0.562	1.000✓	1.000✓
LDR	0.669	0.669	0.674	0.626	0.670	0.720✓
PKN	0.320	1.000✓	1.000✓	0.973	0.608	0.562
GCD	0.997	1.000✓	1.000✓	1.000✓	1.000✓	1.000✓
DBC	0.984	0.998	0.999✓	0.966	0.999✓	0.960
WBC	0.921	0.960	0.970✓	0.946	0.951	0.964
AVG	0.822	0.921	0.923	0.846	0.871	0.858
✓	1	2	4	1	2	3

TABLE III. THE SPECIFICITY OF DCA TOWARDS FEATURE SELECTION

	GR	IG	CS	SVM	SU	GA
BIO	0.966	0.622	0.615	0.707	0.617	0.967✓
IRIS	0.952	0.949	0.953	0.954	0.951	0.992✓
WINE	0.738	0.774	0.769	0.716	0.737	0.839✓
LDR	0.999✓	0.999✓	0.913	0.788	0.999✓	0.986
PKN	0.303	0.339	0.337	0.990	0.067	1.000✓
GCD	0.988✓	0.943	0.945	0.836	0.944	0.953
DBC	0.985✓	0.945	0.951	0.997	0.950	0.900
WBC	1.000✓	1.000✓	1.000✓	1.000✓	1.000✓	1.000✓
AVG	0.866	0.821	0.810	0.873	0.783	0.955
✓	4	2	1	1	1	5

TABLE IV. DCA ACCURATENESS TOWARDS FEATURE SELECTION

	GR	IG	CS	SVM	SU	GA
BIO	0.884✓	0.687	0.680	0.724	0.683	0.884✓
IRIS	0.950	0.946	0.950	0.951	0.949	0.968✓
WINE	0.809	0.834	0.831	0.674	0.808	0.882✓
LDR	0.808	0.808	0.775	0.694	0.808	0.832✓
PKN	0.307	0.502	0.500	0.985✓	0.200	0.669
GCD	0.991✓	0.960	0.962	0.886	0.961	0.967
DBC	0.984	0.964	0.967	0.986✓	0.967	0.921
WBC	0.949	0.974	0.980✓	0.964	0.968	0.976
AVG	0.835	0.834	0.831	0.858	0.793	0.887
✓	2	-	1	2	-	4

In specificity analysis that measures the ability of DCA to detect normal item correctly as normal, all feature selection algorithms demonstrates a similar capability in classifying normal item for WBC, DBC, GCD, LDR, and IRIS dataset with high accuracies. However for BIO and PKN dataset, the GR, IG, CS and SU generate impropriate feature for DCA when their SNS scores are significantly lower than other algorithms. Table III shows the DCA's specificity result. Interestingly, the GA outperforms others algorithm with a high significant result in BIO, IRIS, WINE, and PKN dataset. In comparison to sensitiveness analysis, the GA leads the overall dataset as the best feature selection approach when attains the highest specificity score in 5 datasets (AVG SPS = 0.955). It is followed by GR. In this analysis, the CS algorithm is found to be less performed although it was the best algorithm in the previous sensitivity analysis.

Table IV shows the accurateness of DCA towards different feature selection algorithm. Each algorithm generates a similar result where a comparable result with no significant different are recorded as achieved in specificity analysis. The GA demonstrates the best feature selection algorithm when ranks the highest ACC in 5 datasets out of 8. Besides that, GA also has a consistent result in all dataset when its average score is the highest among other algorithms.

Based on the existing analysis, to decide which approach suggest the most influence features for DCA is a challenging task, especially when the results generated for every evaluation metric are different such that it may have a good score for sensitivity but perform less well in terms of SPS and ACC. In the previous analysis, the chi square is the most well performed approach in term of sensitivity while the rough set genetic algorithm reduct is good at specificity and accuracy. For a model to be a good detection model, it must have the ability to generate a balanced result for sensitivity, specificity, and accuracy [10]. To determine this, the preference matrix approach is implemented in this study [11]. This approach suggests the best model based on the accumulative score of each of the evaluation metrics. Score 1 (the best) until 6 (the worst) is given for the best mining result and the lowest accumulative score indicates the best model. The information in Table V is a summarization of the preference matrix where the total score of every evaluation metrics is depicted in the last row (∑ Score). From the table, the GA has the lowest score thus it is chosen as the most effective feature selection for DCA.

TABLE V. THE PREFERENCE MATRIX

	GR	IG	CS	SVM	SU	GA
SPS	27	18	15	29	17	25
SNS	20	25	26	25	28	15
ACC	25	31	24	28	32	15
∑ Score	72	74	65	82	77	55

VI. CONCLUSION

The most effective technique to select appropriate feature for DCA's signal is investigated in this study. Six feature selection algorithms are experimented and tested on several universal classification data. From the experiment, the rough set genetic algorithm reduct (GA) is found to be the most effect feature selection approach when it generates a consistent result all evaluation metrics. In single evaluation metrics, the chi square has the best competence in term of sensitivity while the GA in both specificity and accuracy. Although GA outperform the other techniques, several issues need to be covered such in selecting only the best suggested feature since it generates all possible features is generated. This is an expensive solution to the problem and is only practical for very simple data sets. In the next step, further analysis will be conducted on complex dataset such as time series data set.

REFERENCES

[1] J. Greensmith, U. Aickelin, and S. Cayzer, "Introducing Dendritic Cells as a Novel Immune Inspired Algorithm for Anomaly Detection " in *4tn International Conference in Artificial Immune Systems (ICARIS)*, 2005, pp. 153-167.

[2] J. Greensmith, "The Dendritic Cell Algorithm," PhD, University of Nottingham, 2007.

[3] F. Gu, J. Greensmith, and U. Aickelin, "Further Exploration of the Dendritic Cell Algorithm: Antigen Multiplier and Time Windows," in *Artificial Immune Systems*. vol. 5132, P. Bentley, D. Lee, and S. Jung, Eds., ed: Springer Berlin Heidelberg, 2008, pp. 142-153.

[4] F. Gu, J. Greensmith, R. Oates, and U. Aickelin, "PCA 4 DCA: The Application Of Principal Component Analysis To The Dendritic Cell Algorithm," in *9th Annual Workshop on Computational Intelligence*, University of Nottingham, 2009.

[5] Z. Chelly and Z. Elouedi, "RC-DCA: A New Feature Selection and Signal Categorization Technique for the Dendritic Cell Algorithm Based on Rough Set Theory," in *Artificial Immune Systems*. vol. 7597, C. Coello Coello, J. Greensmith, N. Krasnogor, P. Liò, G. Nicosia, and M. Pavone, Eds., ed: Springer Berlin Heidelberg, 2012, pp. 152-165.

[6] Z. Chelly and Z. Elouedi, "QR-DCA: A New Rough Data Pre-processing Approach for the Dendritic Cell Algorithm," in *Adaptive and Natural Computing Algorithms*. vol. 7824, M. Tomassini, A. Antonioni, F. Daolio, and P. Buesser, Eds., ed: Springer Berlin Heidelberg, 2013, pp. 140-150.

[7] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, Eds., *Numerical recipes in C*. Cambridge: Cambridge University Press, 1998, p.^pp. Pages.

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

[9] StatLib. (2005, 3 Febuary 2014). *"Statlib — datasets archive"* Available: http://lib.stat.cmu/datasets

[10] M. F. Mohamad Mohsin, A. Abu Bakar, and A. R. Hamdan, "Outbreak detection model based on danger theory," *Applied Soft Computing,* vol. 24, pp. 612-622, 2014.

[11] L. Al Shalabi and Z. Shaaban, "Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix," in *International Conference on Dependability of Computer Systems, 2006. DepCos-RELCOMEX '06.* , 2006, pp. 207-214.

.