Insights into the Antigen Sampling Component of the Dendritic Cell Algorithm

Chris. J. Musselle

Dept. of Computer Science, University of Bristol, UK
chris.musselle@bristol.ac.uk

Abstract. The aim of this paper is to investigate the antigen sampling component of the deterministic version of the dendritic cell algorithm (dDCA). To achieve this, a model is presented, and used to produce synthetic data for two temporal correlation problems. The model itself is designed to simulate a system stochastically switching between a normal and an anomalous state over time. By investigating five parameter values for the dDCA's maximum migration threshold, and benchmarking alongside a minimised version of the dDCA, the effect of sampling using a multi-agent population is explored. Potential sources of error in the dDCA outputs are identified, and related to the duration of the anomalous state in the input data.

1 Introduction

The Dendritic Cell Algorithm (DCA) is an immune inspired algorithm that was developed by Greensmith [1] as part of an interdisciplinary research project between computer scientists and immunologists. The algorithm is an abstract model of dendritic cell behaviour according to the paradigm of danger theory [2], and aims to perform anomaly detection by correlating a series of informative signals (termed either 'danger' or 'safe'), with a sequence of repeating abstract identifiers (termed 'antigens') in the dataset.

The DCA is based on complex processes in the human immune system, however this has resulted in an algorithm with many interacting components and parameters, which has made identification of the key underlying mechanisms a difficult process. This lack of understanding in the factors responsible for its performance, has meant much trial and error in finding suitable parameters when implementing the DCA [3].

A major drawback of the DCA is its reliance on user expert knowledge to define both the parameter values, and also the mapping from the raw data into appropriate inputs for the DCA (antigens and signals). This means that although many applications of the DCA are built on the same underlying framework, the definition of inputs and parameters can be somewhat arbitrary. This is undesirable as user-chosen heuristics are unlikely to give the best results, especially as there is dependency not only between the parameters themselves but also between the parameters and the input data.

A solution to the above shortcomings would be to have an automated process for determining the mapping of the inputs to the DCA, as well as the appropriate parameters. However the relationship between input data and DCA performance is not fully understood, and previous work attempting to address the problem of parameter tuning has proved unsuccessful [4].

Previous work by Stibor et al. [5] has investigated the signal processing side of the algorithm. As a complement to the work of Stibor et al., this current paper aims to explore the antigen sampling side of the algorithm, by designing a model to generate controllable synthetic data for two temporal correlation problems. It is hoped that by examining how the DCA fairs against a selection of different problem scenarios, insight will be gained on how best to frame the problem for the DCA to solve. For the purpose of this paper the deterministic version of the DCA (dDCA) will be used for ease of analysis.

This paper is organised into the following sections. Section 2 covers the background of the DCA, the dDCA and related work. Section 3 outlines the problem statement and proposed solution. Section 4 covers the implementation, with details of the model used to generate the synthetic data, and the experimental design. Section 5 is the evaluation and discussion, with Section 6 stating the conclusions and further work.

2 Background

2.1 Versions of The DCA

The DCA has undergone many revisions since its original inception, resulting in multiple versions of the algorithm in the literature. The two main versions are the original or classic version of the DCA [3], detailed in Greensmiths PhD thesis [1], and the deterministic version of the DCA (dDCA), presented in [6].

The original DCA has in excess of 10 parameters, and employs many stochastic elements, making it exceptionally difficult to formally analyse. Therefore, the deterministic version was formulated to help gain better insight into the underlying mechanisms of the algorithm. The main differences between the two, are that all stochastic components were removed or replaced, and the signal processing calculation was simplified, originally proposed in [4].

Other modified versions have also been presented; which have tended to either add additional components to the algorithm [7] [8], or are based on the underlying framework but differ in their method of implementation [9].

In this study, only the dDCA will be considered, as it represents the simplest version of the algorithm for analytical purposes, and has also shown to give comparable results to the classic DCA [3] [6]. For a full account of the original DCA's inner workings, and the abstractions made from its biological roots, the reader is referred to [10].

2.2 The Deterministic DCA

What follows in this Section, is a detailed description of the components to the dDCA, which can be seen to have three related parts, the signal processing calculation, the antigen sampling mechanism, and calculation of the output metric. Pseudo code for the dDCA can be found in [6].

The inputs to the dDCA take two forms, antigens and signals. Individual antigen are elements of a finite set of integers, of size L, that act as an abstract identifier for some component or event occurring within the monitored system. Examples of such are the process IDs in a computer, the IP addresses in a network, or a node in a sensor network. The number that occur at any one time step is variable, with repeated entries possible.

Each input signal is a time series of length T containing real values, normalised on the interval $[0, 100]$. Two are used in the dDCA, termed the 'safe' and 'danger' signal, and are set to monitor some informative feature of the data determined a priori by expert knowledge of the system in question.

Therefore at each time step $t = \{1, \ldots, T\}$, the inputs to the dDCA consist of the values of the safe and danger signals S_t and D_t, as well as a variable number of antigens A_t. The algorithm itself consists of N virtual dendritic cell (DC) agents, which carry out the signal processing and antigen sampling components.

Signal Processing and Antigen Sampling. All N DCs in the population, sample the same signal values at each time step t, and calculate the following values of

$$csm = S_t + D_t \quad \text{and} \quad k = D_t - 2S_t \ .$$

The variables csm (from the term 'co-stimulatory molecules' [10]) and k are stored internally by each agent $\text{DC}_i = \{1, \ldots, N\}$, as two separate cumulative sums, call them CSM_i and K_i.

The number of antigens present at each time step A_t (termed 'pool'), is assigned to the DC population in a round-robin fashion. That is, each DC agent is assigned one antigen in turn, until all antigens in the pool have been assigned. This process is then repeated with the next pool of antigens in subsequent time steps, however you only give a DC agent a second antigen, when all DCs have received one, and a third after every DC agent has received two, etc.

It is of note that at each time step t, all DCs will process the signals and update their values of CSM_i and K_i, but if $A_t < N$ only a fraction of the DCs will sample additional antigens.

Each DC_i is assigned a different threshold M_i, termed the 'migration' threshold, and once $CSM_i \geq M_i$, DC_i outputs the value of K_i, now termed K_{out}. The time steps up to and including when $CSM_i \geq M_i$ are termed the DCs 'lifetime'.

For all antigens sampled by DC_i during its lifetime, they are tagged as normal if $K_{\text{out}} < 0$ and anomalous if $K_{\text{out}} > 0$. The results of the tagging are logged, and the values of CSM_i and K_i reset to zero. All sampled antigens are also cleared. DC_i then continues to sample signals and collect antigens as before.

Final Classification - the MCAV and K_α Metric. In the original DCA, a final metric, termed the Mature Context Antigen Value or $MCAV$ is calculated for each unique antigen type $l = \{1, \ldots, L\}$ according to

$$MCAV_l = \frac{Anomalous_l}{Total_l} \ ,$$

where $Anomalous_l$ is the number of times antigen l was tagged as anomalous, and $Total_l$ the amount of times antigen l was tagged in total. The $MCAV$ therefore can be thought of as the probability that a given antigen type is anomalous.

This value of the $MCAV$ is then thresholded to achieve the final binary classification of normal or anomalous to each antigen type.

The K_α metric, an alternative metric to the $MCAV$, was proposed with the dDCA in [6]. The K_α uses the average of all output values K_{out} as the metric for each antigen type, instead of first thresholding them at zero into binary tags. The rational behind this is that with K_α the magnitude of K_{out} is taken into account, not just its sign, and therefore may offer more polar separation between antigen types.

However, the use of K_α requires the prior knowledge of all signal values in order to calculate the appropriate threshold of classification, and is also dependent on the weights used to calculate the individual k values.

In the work presented, only the former $MCAV$ metric is used, as it generates a more intuitive output score, capable of being interpreted without the definition of an additional threshold.

2.3 Related Work

In [4] the classic DCA was analysed using frequency analysis, which equated a single DC agent to a combination of filters, specifically those of a moving average filter followed by a downsampler. An attempt was made to automatically tune the M_i distribution using a model representing the transfer function of the combined filters. The results were however unsuccessful, and in an attempt to extend the model to represent multiple DC agents, limitations of the frequency analysis approach, linked to the heterogeneous nature of the DC population were identified [11].

In [5] the authors showed that the signal processing component of the DCA functions as a collection of linear classifiers. Further more, these were also shown to be parallel across the population of DCs, which suggests severe limitations in the regions of signal space the DCA can distinguish between. Across a population of DCs, the positions of the migration decision boundaries were observed to be dependent on past input signal values. This makes the number of DC migrations at any one point in the algorithm very hard to predict.

3 Problem statement and proposed solution

Some of the main barriers to understanding the DCA are the high number of user specified parameters, and the inherent interactions among the algorithm's many components. In addition, the mapping to the inputs for the algorithm must also be user defined, making it difficult to compare the results of DCA applications, as the nature of the inputs will tend to vary in each implementation.

The work of Stibor et al [5] focused solely on the signal processing component of the deterministic DCA, and suggested that more work be done to explore the antigen sampling elements of the algorithm, and its interplay with the signal processing side. This paper sets out to investigate this, through the use of controllable synthetic data, to generate different problem scenarios.

Two experiments will present different types of temporal correlation problems. One will vary the frequency of anomalies in the data set, while the second will investigate how offsetting the antigens and signals affects the outputs of the dDCA. The aim is to highlight which problem scenarios the dDCA performs best at, and what parameter values give these results relative to the input data.

The antigen sampling of the dDCA is controlled by the minimum and maximum values of M_i, which dictate the size of the window (in terms of signal time steps) in which the antigen are correlated with the signals [4]. These will be the primary parameters of interest.

Also, to investigate the effectiveness of population sampling in general, which is employed by the dDCA and all versions of the DCA, a simplified version of the dDCA, termed the minimised dDCA (min-dDCA), will also be implemented as a benchmark. In the min-dDCA, the usual population sampling strategy is replaced by a direct 1-to-1 correlation between signals and antigens.

4 Implementation

This section describes the methodology of the experimental process providing a detailed account of the model used to generate the data and the experimental design for the temporal correlation problems.

Section 4.1 describes in detail the stochastic model used to generate synthetic data for use with the dDCA. The model is designed to simulate a simple system that transitions between a normal and an anomalous state repeatedly over a fixed number of time steps. Data are then generated, taking the form of antigens and signals to be used as inputs for the dDCA. The values of the signals and antigens at each time step are dependent on the underlying state (i.e. normal or anomalous) in a stochastic fashion. The model has three main components which correspond to the generation of the underlying states, the signals, and the antigen pool at each time step in the simulation.

In Section 4.2, the two versions of the dDCA used in this paper are explained, along with the two experiments E1 and E2, that were designed to evaluate their performance.

4.1 Synthetic Data Model

The first component of the model generates a state string q of length n, whose terms $q_t = \{q_1, \ldots, q_n\} \in \{\mathrm{N}, \mathrm{A}\}$ and determine the state of the system at time step t in the simulation ('N' indicates a normal state, and 'A' an anomalous state). To generate q, a simple two state Markov Chain is used, shown in Figure 1a , whose states correspond to normal and anomalous, with transitions between them determined by the probabilities P_{NA} and P_{AN}. The automaton is started

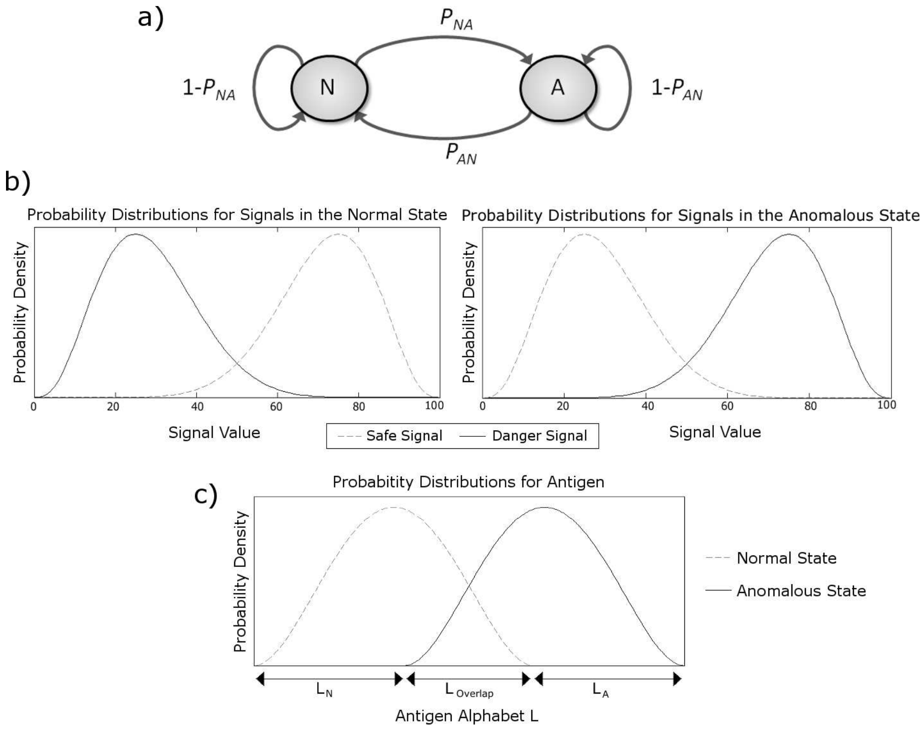

Fig. 1. Overview of the components of the model used to generate synthetic data. Part a) shows the two state Markov Chain used to generate the state string q, part b) the distributions for safe and danger signals in the normal and anomalous state, and part c) the distributions over the antigen alphabet L for each state.

in the normal state and run for n consecutive iterations to generate the output string of states q.

The model then uses q to generate two signal time series $s_t = \{s_1, \ldots, s_n\}$ and $d_t = \{d_1, \ldots, d_n\}$, whose elements are all real valued on the interval $[0, 100]$, and correspond to the levels of danger and safe signals at time t.

These values are generated by drawing from a set of two distinct but overlapping beta distributions, the densities of which are dependant on $q\left(t\right)$. For simplicity, the distributions for s_t and d_t for each state are skewed mirror images of each other, and their positions are exactly opposite for the two separate states of normal and anomalous, as shown in Figure 1b. This results in predominantly low values of danger and high values of safe signal on time steps where the state sequence is the normal state, and vice versa for the anomalous state. The choice of beta distributions was made as they have the useful properties of being bounded and easily tunable to give skewed distributions.

The sequence of antigens A is coded for in a m-by-t matrix with m the number of antigens in the pool per time step t in the simulation. Each pool of antigens $A\left(*, t\right)$, was obtained by drawing values from discreet beta distributions over the finite antigen alphabet L, which dictates all possible antigen types. As with the signal time series, the distribution drawn from is determined by the state at q_t. These distributions, shown in Figure 1c, differ in their range and density over the alphabet L, resulting in two classes of antigen, L_{N} and L_{A}, corresponding to the normal and anomalous state. If these two distributions overlap, then a fraction of the alphabet L_{Overlap} appears under both states.

In summary, four sequences are generated using this model for synthetic data, whereby at time step t in the simulation q_t determines the true state of the system as either normal or anomalous, s_t and d_t are the levels of safe and danger signal respectively, and $A(*, t)$ are the pool of antigens that occur. All parameters used for this model, along with their default values are shown in Table 1.

Table 1. Default Model Parameter Values – Those varied in experiment E1 are annotated with *.

Parameter	Value	Description
Parameters that define state sequence q		
n	500	Number of time steps in simulation.
P_{NA}	0.1*	Probability of transition from normal to anomalous state.
P_{AN}	0.5*	Probability of transition from anomalous to normal state.
Parameters that define signal time series s and d		
$B(\alpha_{\mathrm{SN}}, \beta_{\mathrm{SN}})$	(10, 4)	Beta parameters for s distribution in normal state.
$B(\alpha_{\mathrm{SA}}, \beta_{\mathrm{SA}})$	(4, 10)	Beta parameters for s distribution in anomalous state.
$B(\alpha_{\mathrm{DN}}, \beta_{\mathrm{DN}})$	(4, 10)	Beta parameters for d distribution in normal state.
$B(\alpha_{\mathrm{DA}}, \beta_{\mathrm{DA}})$	(10, 4)	Beta parameters for d distribution in anomalous state.
Parameters that define antigen sequence A		
L	10	Antigen alphabet size.
L_{N}	0.5	Fraction of L that occurs in the normal state.
L_{A}	0.5	Fraction of L that occurs in the anomalous state.
L_{Overlap}	0	Fraction of L that can appear in both states.
m	100	Events per time step.
$B(\alpha_{\mathrm{A}}, \beta_{\mathrm{A}})$	(3, 3)	Beta parameters for the antigen distributions, same for both normal and anomalous state.

4.2 Experimental Design

The research reported here investigates how changing the threshold range across the DC population affects the dDCA output for two temporal correlation problems. Two versions of the dDCA are used in these experiments. The first version is the standard dDCA, with parameters set to the literature values [6], unless stated otherwise. The second is a further simplified version of the dDCA, the min-dDCA.

The min-dDCA is implemented here as a benchmark, to investigate what happens when heterogeneous population sampling, a strategy common to all versions of the DCA, is replaced by a simple 1-to-1 correlation between signals and antigens. In this way it is hoped that insights into how the addition of a population of DCs with variable thresholds impacts on performance, and how best to set the parameters that dictate this variability.

The min-dDCA is a version of the dDCA with only a single DC agent instead of a population of DCs. This one DC performs all the sampling and signal processing throughout the simulation. No migration threshold M is considered, instead the DC will have a lifetime of one, meaning that at every time step in the simulation, the DC will tag all antigens at that time step.

The threshold minimum and maximum for the standard dDCA are set to be related to the average total signal value per time step, $AveTotal$, which can be calculated, as the transition matrix of the states and the signal distributions are known (refer to Table 1 for descriptions of the terms used below).

In order to do so, first the stationary distribution, $\boldsymbol{\pi}$, is calculated from the transition matrix

$$\mathbf{Tr} = \begin{bmatrix} 1 - P_{\text{NA}} & P_{\text{NA}} \\ P_{\text{AN}} & 1 - P_{\text{AN}} \end{bmatrix} ,$$

according to $\boldsymbol{\pi}\mathbf{Tr} = \boldsymbol{\pi}$ [12, p.38]. The stationary distribution gives the ratio A:N which the two state Markov Chain tends to over time. From this, probabilities $P(N)$ and $P(A)$ can be calculated, and used along with the expected values for the Beta distributions to calculate $AveTotal$.

$$E\left[B(\alpha,\beta)\right] = \frac{\alpha}{\alpha+\beta}$$

$$\begin{aligned}
AveSafe &= P(N) \times \frac{\alpha_{\text{SN}}}{\alpha_{\text{SN}}+\beta_{\text{SN}}} + P(A) \times \frac{\alpha_{\text{SA}}}{\alpha_{\text{SA}}+\beta_{\text{SA}}} \\
AveDanger &= P(N) \times \frac{\alpha_{\text{DN}}}{\alpha_{\text{DN}}+\beta_{\text{DN}}} + P(A) \times \frac{\alpha_{\text{DA}}}{\alpha_{\text{DA}}+\beta_{\text{DA}}} \\
AveTotal &= AveSafe + AveDanger \ .
\end{aligned}$$

In the experiments that follow, the migration threshold minimum, M_{MIN} is set to $0.5 \times AveTotal$ and the migration threshold maximum, M_{MAX}, is varied at $1, 1.5, 2, 2.5$ and $3 \times AveTotal$. Therefore six versions of the dDCA were used in the experiments below, the min-dDCA and five versions of the standard dDCA with differing values for the threshold maximum.

All experiments and generation of synthetic data sets were implemented in Matlab version 7.8.0.347 (R2009a) on a 3.16GHz Intel Core2 Duo machine running CentOS Release 5.2.

Due to the stochastic nature of the data sets generated in these experiments, the results from 20 separate initial conditions are reported for each data set.

4.3 The Experiments

Significance of the results between the six different versions of the dDCA were tested for with the null hypothesis $\mathbf{H_0}$; that the results for each version of the dDCA come from the same underlying distribution.

E1: Varying the frequency of the anomalous state. In this series of experiments, the transition probabilities P_{NA} and P_{AN} will be varied from 0.1 to 0.5 to create 25 different datasets on which to test all versions of the dDCA. These probabilities dictate the duration and frequency of each state in the data set generated.

E2: Varying the time delay between signals and antigens. Though an investigation has already been performed into offsetting the signals and antigens in [6], it is reproduced here with varying threshold maxima for the DC population. This experiment therefore investigates which values of the threshold maximum give the best results when a delay of 1 to 5 time steps is introduced between antigens and signals (antigen occurring first). Similarly the experiment also explores a delay of 1 to 5 time steps between the signals and antigens (signals occurring first).

5 Evaluation and Discussion

In assessing the performance of the DCA, the output values of $MCAV$ for these experiments, are not thresholded in the usual way.

Instead a different measure of performance is derived from first averaging the raw values of the $MCAV$ over each antigen class L_{N} and L_{A}. The absolute difference between these averages is then used as a distance metric, $\widehat{D}$. Formally this is defined as,

$$\widehat{D} = \left\| \frac{\sum_{l \in L_{\text{N}}} MCAV_l}{\#L_{\text{N}}} - \frac{\sum_{l \in L_{\text{A}}} MCAV_l}{\#L_{\text{A}}} \right\| ,$$

where $\#L_{\text{N}}$ and $\#L_{\text{A}}$ are the number of unique antigens in class L_{N} and L_{A} respectively. As with the $MCAV$ values themselves, the distance metric $\widehat{D} \in [0, 1]$.

The $\widehat{D}$ metric is a summary statistic which best serves the purposes of this paper, offering a more fine grained assessment of algorithm performance, as the raw output values are used directly, and not thresholded to a binary classification. It also represents a common way of assessing all experimental results equally, without the need to assign an additional threshold for the algorithm.

5.1 E1: Varying the Frequency of the Anomalous State

Varying the value of P_{NA} between 0.1 and 0.5 showed no significant difference between the $\widehat{D}$ metrics, for all six versions of the dDCA. This was tested for using the Kruskal-Wallis statistical test (results not shown). However, varying

P_{AN} between 0.1 and 0.5 causes a notable decrease in the $\widehat{D}$ metric for all versions of the standard dDCA. The larger the maximum threshold M_{MAX}, the greater the decrease in $\widehat{D}$, as shown in Figure 2. The min-dDCA however was observed to have a steady value of $\widehat{D}$ as P_{AN} was increased. Only at $P_{\mathrm{AN}} = 0.1$, was the null hypothesis $\mathbf{H_0}$ not rejected, for all other data sets it was successfully rejected with $p < 0.01$.

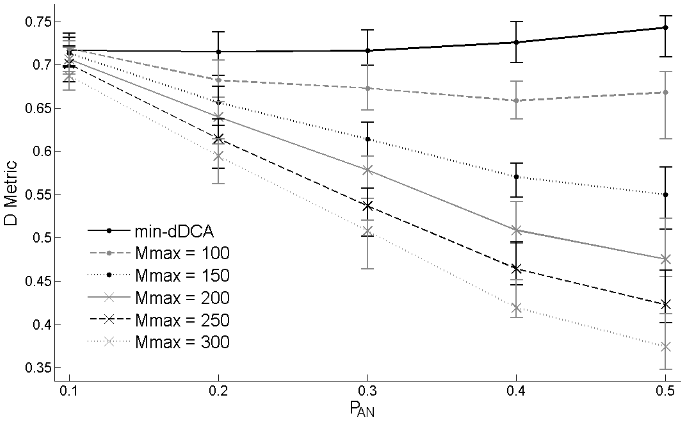

Fig. 2. Change in the distance metric D with transition probability P_{AN} for all six versions of the dDCA. The spread of all 20 $\widehat{D}$ values are shown at each data point, with the point plotted at the median, and the error bars corresponding to the 25th 75th percentile. Null hypothesis $\mathbf{H_0}$ was rejected with $p < 0.01$ when $P_{\mathrm{AN}} \geq 0.2$

The relationship between $\widehat{D}$ and M_{MAX} is explained by the fact that larger values of M_{MAX} will mean more DCs in the population having longer sampling lifetimes. This equates to more DCs in the population sampling wider time windows of signals, along with the associated antigens. Some of these time windows will fall across the boundaries between the normal and anomalous state in the signal time series, and as antigens of both classes will be collected during this time window, misclassification of some antigens is inevitable. In fact, due to the heavy bias for the safe signal when calculating k (see Section 2.2), sampling across such boundaries is highly likely to tag all antigens collected as normal. The frequency of such misclassification increases as M_{MAX} is increased, lowering the $MCAV$ for the anomalous antigen class and therefore the observed difference between the classes, $\widehat{D}$.

The decrease in $\widehat{D}$ with P_{AN} is related to the above. Higher values of P_{AN} will give rise to shorter consecutive anomalous states in the data set, and lead to more transition boundaries between the normal and anomalous state in the time series. This will cause a higher proportion of the anomalous antigens to occur at the boundaries, and as the antigens occurring at these points are prone to misclassification by DCs with longer life times (larger sampling time windows), the metric $\widehat{D}$ decreases with increasing P_{AN}.

Increasing P_{NA} has little impact on $\widehat{D}$, as it dictated the level of consecutive normal states in the data set. Due to the high bias in calculating k, the normal antigens at the boundaries are almost always tagged correctly, meaning that changing the proportion of antigens occurring at these boundaries has little effect.

The min-dDCA performs direct 1-to-1 correlation between signals and antigens throughout the experiment. While this approach was observed to be the best for this problem scenario, it is noted that the temporal correlation required is trivial in nature, and so the capabilities of the dDCA may not have been realised in such a simple task. The experiment does however highlight for the standard dDCA, a relationship between the duration of the anomalous state in the data set, and the average difference between the $MCAV$ values for each antigen class, as measured with $\widehat{D}$.

5.2 E2: Time-Offset Between Antigen and Signals

In E2, the time-offset between antigens and signals was varied between -5, (antigens appearing before signals) to +5 time steps (antigens appearing after signals). The experiment was repeated with values of P_{AN} at 0.1, 0.3 and 0.5, which are shown in Figures 3a, 3b and 3c respectively. This was done in light of the results of E1, as the duration of the anomalous state was seen to affect the output values of $\widehat{D}$ considerably.

The results show a general trend of decreasing $\widehat{D}$ when the time-offset is increased in either direction form the origin. The rate of this drop off is increased when P_{AN} is increased, and also appears symmetric about the origin.

The overall decrease in $\widehat{D}$ with increasing time-offset is expected as the corresponding antigen pool becomes increasingly out of phase with the signals set to indicate its state, leading to greater misclassification errors for antigens of both classes. However this effect is amplified when P_{AN} is increased, as when the number of consecutive anomalous states is reduced, there will be less overlap between the out of phase anomalous antigens and the underlying states.

Significant differences between the $\widehat{D}$ values for the various dDCA versions were tested for using the Kruskal-Wallis statistical test. In Figure 3, those annotated with * indicate successful rejection of the null hypothesis $\mathbf{H_0}$ with $p < 0.05$ and ** with $p < 0.01$.

For $P_{\text{AN}} = 0.1$ in Figure 3a these differences were most notable between the min-dDCA and versions with a higher value of M_{MAX}. At time-offsets of 1/-1 and 2/-2 the standard dDCA with M_{MAX} in the range of 200-300 showed greater values of $\widehat{D}$ when compared to the min-dDCA. For $P_{\text{AN}} = 0.3$ in Figure 3b however, no such trend is noted, and in Figure 3c, this trend is reversed, with the min-dDCA having greater values of $\widehat{D}$ when compared to the standard dDCA with larger values of M_{MAX}.

It seems that when the average duration of the anomalous state is sufficiently high, achieved by a lower value of P_{AN} for the data set, sampling over longer time windows is preferable over a direct 1-to-1 correlation, as this has the effect of mitigating errors introduced by the signals and antigens being out of phase by 1-2 time steps. This potentially beneficial effect exhibited by the standard dDCA

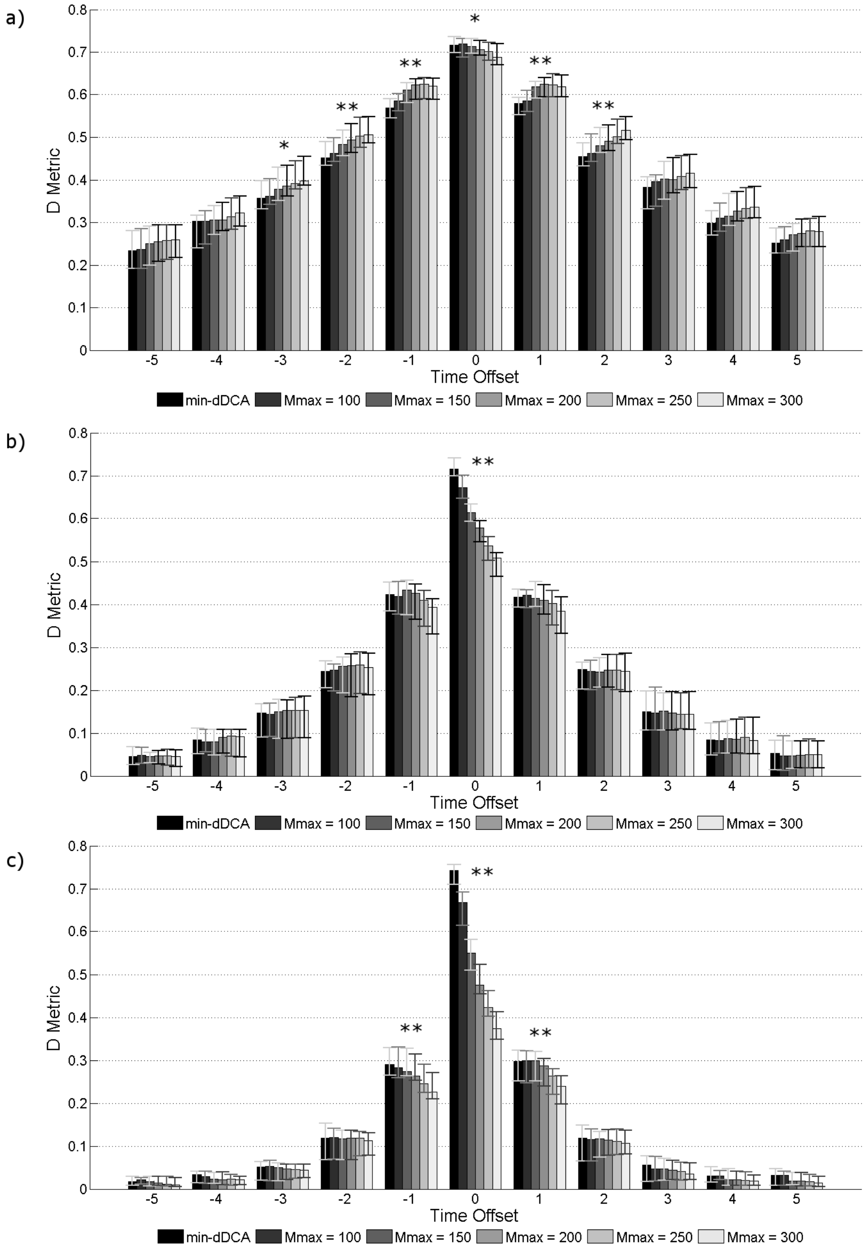

Fig. 3. Change in the distance metric $\widehat{D}$ when varying the time-offset between -5, (antigens appearing before signals) to +5 time steps (antigens appearing after signals). The experiment was repeated with values of P_{AN} at 0.1, 0.3 and 0.5, which are shown in Figures a), b) and c) respectively for all six versions of the dDCA. The spread of all 20 $\widehat{D}$ values are shown at each data point, with the bar plotted at the median, and the error bars corresponding to the 25th and 75th percentile. A Kruskal-Wallis statistical test was performed to test for a difference in the values of $\widehat{D}$ across the various dDCA versions, those annotated with * indicate $p < 0.05$ and ** $p < 0.01$

is however reversed when the number of consecutive anomalous states is reduced (higher values of P_{AN}).

6 Summary and Conclusions

This work introduces a novel model to artificially generate input data for use with the dDCA. Data sets for two experiments were generated using this model, one investigating the effects of altering the frequency and duration of anomalies in the input data, and the second investigating the effect of introducing a delay between signals and antigens (and vice versa). The performances of the six different versions of the dDCA tested, were defined in terms of a distance metric $\widehat{D}$ detailed in Section 5, measuring the average difference between $MCAV$ values for the two antigen classes.

Form the first experiment, it can be concluded that an important factor to consider for dDCA performance, is the likely duration of the anomalous states in the input data. This is because, due to the high bias of the safe signal, anomalous antigens are often misclassified at the transition boundaries between normal and anomalous states. A shorter duration will mean a greater proportion of the anomalous antigen occurring at these boundaries, lowering the output value for the anomalous $MCAV$'s, and decreasing the difference between the two antigen classes.

This situation may be improved in one of two ways. First the weights used could be changed, so there is no longer such a strong bias for the safe signal. Secondly, segmentation of the input data, originally investigated in [8] to provide a more real-time implementation of the algorithm, could be used to reduce the number of transition boundaries processed at once. Further research would be needed to verify whether this is indeed the case.

The second experiment demonstrated that, for a delay of 1-2 time steps between signals and antigen, the standard dDCAs with higher values of M_{MAX} were able to mitigate the errors introduced by the signals and antigen being out of phase. Further work will be needed to investigate why this is the case, though it is speculated that DCs with longer sampling windows make fewer errors overall in this scenario, as long as the duration of the anomalous state is not too short lived.

The results of the min-dDCA and standard dDCAs are very comparable for the data sets investigated in this work. However, it is noted that the noise introduced to the signal values by the stochastic nature of the model, was not sufficient to cause misclassification errors with any great frequency. Future work will address this issue by incorporating more noise into the generated data sets so they are closer to real world data. This does suggest however that multi-agent sampling is not always necessary to achieve adequate separation of $MCAV$ values between antigen classes, so long as noise in the signal values is kept to a minimum. One way of achieving this would be through appropriate pre-processing of the signal values.

A natural extension of this work, would be to investigate other temporal correlation problems in the same way presented here, by tuning the synthetic model to generate the appropriate input data. In doing so, it may also be appropriate to add extra components to the model, so as to represent more complex data sets. One such extension would be to allow the size of the antigen pool to vary at each time step, which was kept constant in these experiments, to focus primarily on the effect of changing the M_i distribution.

Acknowledgements The author would like to thank Professor Dave Cliff and Dr. Ayalvadi Ganesh for their helpful feedback and advice on the work presented, and the Bristol Centre for Complexity Science (BCCS) for providing funding under the EPSRC grant EP/5011214.

References

1. Greensmith J.: The Dendritic Cell Algorithm, PhD Thesis, The University of Nottingham (2007).
2. Aickelin U., Bentley P., Cayzer S. et al.: Danger Theory: The Link between AIS and IDS? In: Timmis, J., Bentley, P., Hart, E., (eds.) ICARIS 2003. LNCS, vol 2787, pp. 147-155. Springer, Heidelberg (2003).
3. Manzoor S., Shafiq, M.Z., Tabish, S.M., et al.: A Sense of Danger for Windows Processes. In: Andrews, P.S., Timmis, J., Owens, N.D.L., Aickelin, U., Hart, E., Hone, A., Tyrrell, A. (eds.) ICARIS 2009. LNCS vol 5666, pp. 220-233 Springer, Heidelberg (2009).
4. Oates R., Kendall G., and Garibaldi J.: Frequency Analysis for Dendritic Cell Population Tuning: Decimating the Dendritic Cell. Evolutionary Intelligence vol 1, pp. 145-157 (2008)
5. Stibor T., Oates R., Kendall G. et al.: Geometrical insights into the dendritic cell algorithm. In: Proceedings of the 11th Annual conference on Genetic and Evolutionary Computation, pp. 1275-1282 (2009).
6. Greensmith J. and Aickelin U.: The Deterministic Dendritic Cell Algorithm, In: Bentley, P., Lee, D., Jung, S. (eds.) ICARIS 2008. LNCS vol 5132, pp. 291-302. Springer, Heidelberg (2008).
7. Gu F., Greensmith J., and Aickelin U.: Further Exploration of the Dendritic Cell Algorithm: Antigen Multiplier and Time Windows, In: Bentley, P., Lee, D., Jung, S. (eds.) ICARIS 2008. LNCS vol 5132, pp. 142-153. Springer, Heidelberg (2008).
8. Gu F., Greensmith J., and Aickelin U.: Integrating real-time analysis with the dendritic cell algorithm through segmentation, In: Proceedings of the 11th Annual conference on Genetic and evolutionary computation, pp. 1203-1210, (2009).
9. M. Mokhtar, J. Timmis and A. Tyrrell: A modified dendritic cell algorithm for online error detection in robotic systems, In: Proceedings of the Eleventh conference on Congress on Evolutionary Computation, pp. 2055-2062, (2009).
10. Greensmith J., Aickelin U., and Cayzer S.: Detecting Danger: The Dendritic Cell Algorithm. In: Robust Intelligent Systems, 89-112, (2009).
11. Oates R., Kendall G., and Garibaldi J.: The Limitations of Frequency Analysis for Dendritic Cell Population Modelling. In: Bentley, P., Lee, D., Jung, S. (eds.) ICARIS 2008. LNCS vol 5132, pp. 328-339. Springer, Heidelberg (2008).
12. Daniel W. Stroock: An Introduction to Markov Processes, 1st ed. Springer, (2005).