# A New Data Pre-processing Approach for the Dendritic Cell Algorithm Based on Fuzzy Rough Set Theory

Zeineb Chelly*
Department of Computer Science, University of Tunis, Institut Supérieur de Gestion of Tunis, 2000 Le Bardo, Tunis, Tunisia
zeinebchelly@yahoo.fr

Zied Elouedi
Department of Computer Science, University of Tunis, Institut Supérieur de Gestion of Tunis, 2000 Le Bardo, Tunis, Tunisia
zied.elouedi@gmx.fr

## ABSTRACT

The aim of this paper is to develop a new data pre-processing method for the dendritic cell algorithm (DCA) based on Fuzzy Rough Set Theory (FRST). In this new fuzzy-rough model, the data pre-processing phase is based on the fuzzy positive region and the fuzzy dependency degree concepts. Results show that applying FRST is more convenient for the DCA data pre-processing phase yielding much better performance in terms of accuracy.

## Categories and Subject Descriptors

I.2 [**Computing Methodologies**]: Artificial intelligence

## General Terms

Algorithms, Experimentation, Performance.

## Keywords

Artificial immune systems, Dendritic cells, Fuzzy-rough sets, Feature selection.

## 1. INTRODUCTION

One of the emerging algorithms within the field of artificial immune systems is the Dendritic Cell Algorithm (DCA) [1]. Since its commencement, DCA has been based on several modifications especially modifications investigating the algorithm pre-processing phase. Recently in [2], a new DCA approach, named RC-DCA, was introduced; based on Rough Set Theory (RST) [3]. However, the application of RST as a pre-processor is reliant upon crisp datasets; the attribute values of the input database should be discretized beforehand. Consequently, important information may be lost as a result of quantization. Therefore and to avoid this problem, in this paper, we propose to develop a novel DCA method, named FRST-DCA, based Fuzzy Rough Set Theory (FRST) [4] for data pre-processing.

## 2. FRST-DCA: THE SOLUTION APPROACH

FRST-DCA is based on the Fuzzy-Rough QuickReduct algorithm. We focus, mainly, on our FRST-DCA data pre-processing step as the rest of the algorithm steps are performed the same as DCA [1].

**1) Signal Selection Process:** Our problem is seen as an information table, where universe $U = \{x_1, x_2, \ldots, x_N\}$ is a set of antigen identifiers, the conditional attribute set $C = \{c_1, c_2, \ldots, c_A\}$ contains each feature to select and the decision attribute $D$ corresponds to the class label of each sample; having binary values $d_k$: either the antigen is classified as normal or as anomalous. For feature selection, FRST-DCA computes, first of all, the fuzzy lower approximations of the two decision concepts $d_k$, for all attributes $c_i$ and for all objects $x_j$; denoted by $\mu_{c_i\{d_k\}}(x_j)$. Using these results, FRST-DCA calculates the fuzzy positive regions for all $c_i$, for each object $x_j$, defined as $\mu_{POS_{c_i}(D)}(x_j)$. To find the reduct, FRST-DCA starts off with an empty set and moves to calculate the fuzzy dependency degrees of $D$ on $c_i$, defines as $\gamma'_{c_i}(D)$. The attribute $c_m$ having the greatest value of fuzzy-rough dependency degree is added to the empty reduct set. Once the first attribute $c_m$ is selected, FRST-DCA adds, in turn, one attribute to the selected first attribute and computes the fuzzy-rough dependency degree of each obtained attributes couple $\gamma'_{\{c_m,c_i\}}(D)$. The algorithm chooses the couple having the greatest fuzzy-rough dependency degree. The process of adding each time one attribute to the subset of the selected features continues until no increase in the fuzzy-rough dependency degree is observed.

**2) Signal Categorization Process:** Based on the semantics of the DCA signals, SS, PAMP and DS, and since FRST-DCA has already calculated the fuzzy-rough dependency degree of each attribute $c_i$ a part, $\gamma'_{c_i}(D)$, FRST-DCA selects the first attribute $c_m$ having the greatest fuzzy-rough dependency degree to form the SS as it is considered the most informative first feature added to the reduct. With no additional computations and since FRST-DCA has already computed the fuzzy-rough dependency degrees of each attributes couple $\gamma'_{\{c_m,c_i\}}(D)$ when adding, in turn, one attribute $c_i$ to the selected first attribute $c_m$ that represents the SS, FRST-DCA chooses the couple having the greatest dependency degree. More precisely, FRST-DCA selects that second attribute $c_r$ having the greatest $\gamma'_{\{c_m,c_r\}}(D)$ among the calculated $\gamma'_{\{c_m,c_i\}}(D)$; to form the PAMP signal. Finally, the rest of the reduct attributes are combined and affected to represent the DS.

*Corresponding Author


*GECCO'13 Companion,* July 6–10, 2013, Amsterdam, The Netherlands.
ACM 978-1-4503-1964-5/13/07.

**Table 2: Comparison Results of DCA Approaches**

| Database | Specificity(%) | | Sensitivity(%) | | Accuracy(%) | | Time(s) | | ♯ Attributes | |
|---|---|---|---|---|---|---|---|---|---|---|
| | DCA | | DCA | | DCA | | DCA | | DCA | |
| | RC | FRST | RC | FRST | RC | FRST | RC | FRST | RC | FRST |
| SN | 93.82 | 95.88 | 90.10 | 94.60 | 91.82 | 95.19 | 1705.79 | 95.91 | 20 | 16 |
| Bio | 79.24 | 86.79 | 77.35 | 84.90 | 78.30 | 85.84 | 1679.53 | 47.29 | 19 | 13 |
| SP | 98.49 | 99.89 | 98.40 | 99.77 | 98.45 | 99.84 | 3184.83 | 2452.25 | 8 | 8 |
| CylB | 97.75 | 98.71 | 97.00 | 98.00 | 97.46 | 98.43 | 1441.93 | 118.41 | 7 | 7 |
| Ch | 98.88 | 98.69 | 98.80 | 99.28 | 98.84 | 98.99 | 1779.83 | 1047.25 | 11 | 4 |

**Table 1: Description of Databases**

| Database | Ref | ♯ Instances | ♯ Attributes |
|---|---|---|---|
| Sonor | SN | 208 | 61 |
| Molecular-Bio | Bio | 106 | 59 |
| Spambase | SP | 4601 | 58 |
| Cylinder Bands | CylB | 540 | 40 |
| Chess | Ch | 3196 | 37 |

# 3. RESULTS AND DISCUSSION

Our experiments are performed on two-class real-valued attributes databases from [5], described in Table 1. For the DCA approaches, namely FRST-DCA and RC-DCA, each data item is mapped as an antigen, with the value of the antigen equal to the data ID of the item. A population of 100 cells is used. The migration threshold of an individual DC is set to 10. To perform anomaly detection, a threshold which is automatically generated from the data is applied to the Mature Context Antigen Values (MCAVs). The MCAV threshold is derived from the proportion of anomalous data instances of the whole dataset. Items below the threshold are classified as class one and above as class two. The resulting classified antigens are compared to the labels given in the original datasets. For each experiment, the results presented are based on mean MCAV values generated across a 10-fold cross validation. Let us remind that in [2], results showed that applying RST with DCA is a good alternative leading to a better classification performance. However, the developed RC-DCA rough model suffers from a main limitation which is the time taken by the algorithm to process which contradicts the main characteristic of the standard DCA: its lightweight in terms of running time [6]. This limitation is due to the set of all possible reducts generated by RC-DCA. Thus, we have developed the FRST-DCA fuzzy-rough approach. We aim to show that applying FRST, instead of RST, can avoid the information loss caused by the RST mandatory step of data quantization, beforehand. In addition, we aim to show that by leaving the attributes values unchanged, FRST-DCA is able to select fewer features than RC-DCA, leading to better guide the FRST-DCA algorithm classification process. This is confirmed by the results of Table 2. For instance, from Table 2, we can notice that our FRST-DCA, has fewer features than the crisp rough DCA models, RC-DCA. This is explained by the fact that FRST-DCA, by applying the Fuzzy-Rough QuickReduct algorithm, incorporates the information usually lost in crisp discretization by utilizing the generated fuzzy-rough sets to provide a more informed technique. The results show that FRST-DCA selects features without much loss in information content. Our FRST-DCA new approach performs much better than traditional RST on the whole, in terms of both feature selection and classification quality. For instance, applying FRST-DCA to the Bio database, the number of selected attributes is 13. However, when applying RC-DCA to the same database, the number of selected features is set to 19. Furthermore, from Table 2, we can notice that FRST-DCA outperforms RC-DCA in terms of classification accuracy. For instance, when applying the algorithms to the SN dataset, the classification accuracy of FRST-DCA is set to 95.19%. However, when applying RC-DCA to the same database, the accuracy is set to 91.82%. Same remark is observed for the specificity and the sensitivity criteria. When comparing the results in terms of running time, we can notice that the time taken by FRST-DCA to process is less than the time needed by RC-DCA to function as this latter model generates all the set of possible reducts. For example, when applying the algorithms to the Bio database, the amount of time taken by FRST-DCA to process is 47.29(s) which is less than the time taken by RC-DCA which is 1679.53(s).

# 4. CONCLUSION

In this paper, we have proposed a fuzzy rough DCA model for data pre-processing named FRST-DCA. We have shown that FRST-DCA has the advantages of selecting fewer features than the crisp rough DCA developed model; RC-DCA. FRST-DCA is capable of avoiding the information loss caused by the use of the crisp rough DCA model. The application of FRST-DCA to the unchanged attributes values led our new model to better guide its classification task.

# 5. REFERENCES


[1] J. Greensmith, U. Aickelin, and S. Cayzer. Introducing dendritic cells as a novel immune-inspired algorithm for anomaly detection. In *ICARIS*, pages 153–167, 2005.

[2] 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 *ICARIS*, pages 152–165, 2012.

[3] Z. Pawlak. Rough sets. *International Journal of Computer and Information Science*, 11:341–356, 1982.

[4] D. Dubois and H. Prade. *Putting rough sets and fuzzy sets together*. Kluwer Academic Publishers, Dordrecht, 1992.

[5] A. Asuncion and D.J. Newman. UCI machine learning repository, http://mlearn.ics.uci.edu/mlrepository.html, 2007.

[6] J. Greensmith and U. Aickelin. The deterministic dendritic cell algorithm. In *ICARIS*, pages 291–302, 2008.