RC-DCA: A New Feature Selection and Signal Categorization Technique for the Dendritic Cell Algorithm Based on Rough Set Theory

Zeineb Chelly and Zied Elouedi

LARODEC, Université de Tunis, Institut Supérieur de Gestion de Tunis, 41 Avenue de la liberté, cité Bouchoucha, 2000 Le Bardo, Tunisia
zeinebchelly@yahoo.fr, zied.elouedi@gmx.fr

Abstract. The Dendritic Cell Algorithm (DCA) is an immune inspired algorithm based on the behavior of dendritic cells. The performance of DCA depends on the selected features and their categorization to their specific signal types, during pre-processing. For feature selection, DCA applies the Principal Component Analysis (PCA). Nevertheless, PCA does not guarantee that the selected first principal components will be the most adequate for classification. Furthermore, the DCA categorization process is based on the PCA attributes’ ranking in terms on variability. However, this categorization process could not be considered as a coherent assignment procedure. Thus, the aim of this paper is to develop a new DCA feature selection and categorization method based on Rough Set Theory (RST). In this model, the selection and the categorization processes are based on the RST CORE and REDUCT concepts. Results show that applying RST, instead of PCA, to DCA is more convenient for data pre-processing yielding much better performance in terms of accuracy.

Keywords: Dendritic Cells, Principal Component Analysis, Rough Set Theory, Core, Reduct.

1 Introduction

The Dendritic Cell Algorithm (DCA) is an immune inspired algorithm developed in [1] as a part of an interdisciplinary research project between computer scientists and immunologists. DCA is derived from behavioral models of natural dendritic cells (DCs). The DCA has the ability to combine a series of informative signals with a sequence of repeating abstract identifiers termed “antigens” to perform anomaly detection. To achieve this and through the pre-processing phase, the DCA selects a subset of features and assigns each one of them to its specific signal type. Each selected feature can be categorized either as Danger Signal (DS), Safe Signal (SS) or as Pathogen-Associated Molecular Pattern (PAMP). DCA combines these signals internally to produce a set of output signals in combination with location markers in the form of antigen to process his
classification task. According to this procedure, it is clearly noticed that the DCA classification task - the determination of the antigen context - is very dependent on its pre-processing phase which is based on signal selection and signal categorization. More precisely, the selected subset of features and their categorization to definite signal types can influence the DCA classification results.

As the DCA data pre-processing phase would better be automatic to avoid the influence of any external information given by users or experts, in [2], the Principal Component Analysis (PCA) statistical method [3] was introduced in the DCA data pre-processing phase. The use of PCA aims to automatically select the features to retain for the DCA and to perform their categorization to their specific signal types. More precisely and for signal selection, PCA transforms a finite number of possibly correlated vectors into a smaller number of uncorrelated vectors, termed “principal components” which reveals the internal structure of the given data with the focus on data variance [2]. Nevertheless, using PCA as a dimensionality reduction technique presents some shortcomings as it is not necessarily true that the first selected principal components that capture most of the variance are the adequate features to retain [4]. Consequently, choosing these components for the DCA may influence its classification phase by producing unreliable results. Other dimensionality reduction techniques were defined in [5] including correlation coefficient and information gain.

As for feature categorization, DCA uses the PCA ranking of attributes which is based on variability and maps this obtained order to the ranking of the signal categories of the DCA which is in the order Safe, PAMP and Danger implying the significance of each signal category to the signal transformation of the DCA [2]. However, this categorization reasoning which is based on attributes’ ranking and where the variability of attributes is equivalent to importance could not be considered as a coherent and consistent categorization procedure. In [5], the signal categorization step is based on the use of the Mean Squared Error (MSE) [6] which is the negative of the expected value of one specific utility function known the quadratic utility function. However, the quadratic utility function may not be the appropriate function to use under a given set of circumstances. Moreover, MSE has the disadvantage of heavily weighting outliers [7]. This is a result of the squaring of each term, which effectively weights large errors more heavily than small ones. Therefore and by applying these methods, the signal categorization step presents a shortcoming that can influence negatively the DCA functioning. Therefore, in this paper, we propose to develop a novel bio-inspired model of the DCA based on a new signal selection and categorization technique. Our new model is derived from the behavioral models of natural dendritic cells and grounded on the framework of Rough Set Theory (RST) for signal selection and signal categorization.

Rough Set Theory (RST) [8] has been employed with much success in different fields as a dimensionality reduction technique [9,10]. It has also been applied to improve the performance and the classification effectiveness of several algorithms [11]. To select features, RST removes the unnecessary attributes while keeping only the most informative ones - a subset termed REDUCT - that preserve nearly