A Hybrid Approach for Gene Expression Data Clustering

George Barreto Bezerra & Leandro Nunes de Castro

{bezerra,lnunes}@dca.fee.unicamp.br

ABSTRACT

This work proposes a new approach for gene expression data clustering. The technique proposed is based on a combination of two algorithms – aiNet and the minimal spanning tree (MST) – through a complementary hybrid analysis. The aiNet (Artificial Immune NETwork) [1,2] is an artificial immune system inspired by the immune network theory, originally proposed by Niels Jerne (1974) [4]. It is an iterative clustering algorithm that performs data compression using a pattern recognition process inspired by the human immune system. In the biological system, specific antibodies are produced to recognize disease-causing agents, broadly named antigens. In the present work, the "antigens" correspond to the gene expression data, and the antibodies are the aiNet cells, which will be representative of the gene expression data. The aiNet algorithm thus plays two important roles. First, it simplifies the complexity of the problem by compressing the input data set, filtering out outliers and using multiple prototypes for representing different classes of data. Second, aiNet places the prototypes (network cells or antibodies) in regions of the input space relevant for the clustering of multivariate data in spaces of very high dimension, such as gene expression data.

The next step in the analysis is to use the minimal spanning tree to detect inherent separations between the subsets present in the spatial distribution of the network of antibodies. The MST is a tool from graph theory that proved to be a powerful artifice for data clustering [6]. Roughly, given a set of points (data), a MST is built linking all these points, and those links considered inconsistent are removed from the tree, resulting in a disconnected graph. Each of the subgraphs generated correspond to one cluster. There are several forms of evaluating the inconsistency of edges in an MST. It has even already been used for gene expression data analysis [5], but with methods of identification and removal of inconsistent edges completely different from the method used here, which is based on a local criterion originally proposed by Zahn (1971) [6]. In our approach, it is possible to explore cluster boundaries by taking into account their relative densities, thus preserving the inherent structure of the data spatial distribution. Another important aspect of our proposal is that the MST is built on the antibody network, and not directly on the data set. This characteristic has a major impact in the clustering process, because the data compression performed by the aiNet and the prototype positioning reduces the levels of noise and redundancy in the data set and discovers key portions of the input space for the detection and representation of clusters. Therefore, the aiNet makes it possible for the MST to detect inherent separations within the data set.

The hybrid method proposed was applied to a benchmark data set of the yeast *Saccharomyces cerevisiae* gene expression levels obtained in [3]. Four clusters previously detected in [3] were chosen for the analysis: clusters C, E, F and H, totalizing 68 genes in 79 different experimental conditions. These clusters were the same used in the analysis performed in [5]. Using the correlation coefficient as distance measure, the hybrid algorithm was capable of detecting correctly the four clusters in ten test cases, with an average data compression of 21%. Building the MST directly on the data set and applying the proposed inconsistency criterion, the clusters E, F and H were perfectly identified, but cluster C was divided into two subsets. This may be due to the absence of the robustness introduced by the aiNet, making the MST susceptible to noise. This result, together with other empirical investigations currently being performed, suggest that aiNet plays a key role in the detection of important portions of the input space, and thus to be used in the clustering process.

The results also demonstrate the feasibility of the proposed method as a clustering technique. It is capable of accurately detecting the presence of clusters within the data sets studied. Another remarkable advantage of this algorithm is that no knowledge about the number of clusters is required a priori, as the most classical approaches do, such as the hierarchical clustering techniques [3].

REFERENCES

[1] de Castro, L. N. & Von Zuben, F. J. (2001), "aiNet: An artificial Immune Network for Data Analysis", In *Data Mining: A Heuristic Approach*, H. A. Abbass, R. A. Saker, and C. S. Newton (Eds.), Idea Group Publishing, USA, Chapter XII, pp. 231-259.

[2] de Castro, L. N. & Von Zuben, F. J. (2000), "An Evolutionary Immune Network for Data Clustering", *Proc. do IEEE SBRN*, pp. 84-89.

[3] Eisen, M. B., Spellman, P. T., Brow, P. O., & Botstein, D. (1998), "Cluster Analysis and Display of Genome-wide Expression Patterns", Proc. Natl. Acad. Sci, Vol.95, pp. 14863-14868, USA.

[4] Jerne, N. K. "Towards a Network Theory of the Immune System", Ann. Immunol. (Inst. Pasteur), 1974, pp. 373-389.

[5] Xu, Y., Olman, V. & Xu, Dong (2002), "Minimum Spanning Trees for Gene Expression Data Clustering", Bioinformatics, vol. 18, pp. 536-545.

[6] Zahn, C. T. (1971), "Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters", IEEE Trans. on Computers, C-20(1), pp.68-86.