Decision Trees and Random Forests Modeling by using P Systems (Extended Abstract)

José M. Sempere

Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València,
jsempere@dsic.upv.es

Keywords: Decision trees and random forests, P systems with active membranes and communication rules, tissue like P systems, ensemble machine learning, P systems working in an entropic manner.

Decision trees are tree-structured classification models that have been widely used in different application domains such as bioinformatics, pattern recognition, data mining, and so on [4]. The definition of P systems to model decision trees has been previously approached in different works. For example, Díaz-Pernil *et al.* [2] proposed recognizer P systems to define decision trees. Their proposal is based on a non-deterministic search for structures compatible with examples of the classification that should be carried out by the decision tree. Wang *et al* [6] proposed the use of tissue-like P systems with tree-like objects and evolutionary strategies in order to explore a searching space by using the nondeterminism of the P systems. In our approach, we model decision tree by using basic concepts provided by cell-like P systems: the tree structure is defined in a very easy way through the tree-like structure of the regions in a cell-like P system, the rules of the decision tree can be defined by communication rules in the P system and, finally, the selection of the attributes that best fit the learning examples is done using a criterion based on information theory as it happens in the learning algorithms of decision trees. Therefore, we believe that our proposal is better adapted to the use of P systems in a more natural way than the proposals referred to above.

In this work, we provide a method of obtaining a P system with active membranes and communication rules from the definition of a decision tree. This will be done through an algorithmic scheme. From that definition, we provide a method of classifying objects that allows their classification one by one. Also, the classification can be carried out in parallel in a massive way by applying the maximum parallelism of the P systems and a replication of the rules.

The other aspect to consider when working with decision trees is that of their efficient machine learning from examples. Over time, various learning algorithms on decision trees have been proposed. From our point of view, the best known algorithms would be Quinlan's ID3 and C4.5 [3] and the CART algorithm for classification and regression trees [1]. Both proposals are based on the choice of attributes for the construction of the tree based on criteria of information

theory or, more directly, on entropy. According to this criterion, we propose a method of construction of P systems based on systems that work in an *entropic manner* [5]. For this, the method uses membrane creation and communication rules chosen under a criterion of minimization of *self-referred entropy*.

The other concept that we propose in this paper is the definition of random forests using P systems. The random forests are classification methods that allow us to tackle diverse tasks, where the objects to be classified have a high number of attributes and it is not known with certainty which of them are more useful than the others. Therefore, by means of precise learning methods, such as the bootstrap aggregating (*bagging*), several decision trees are constructed and new objects are classified by combining the results of each decision tree, for example through a majority decision. These types of techniques are known as ensemble machine learning [7].

Our proposal for the definition of random forests using P systems, is the use of tissue systems with a star connection topology. In our proposal, each decision tree is defined in a peripheral cell of the tissue, and collection decisions are made at the central node based on pre-established communication rules. We also propose the possibility of initially creating each decision tree in a non-deterministic way. In this case, the attributes of each tree, imitating the bagging techniques, are selected non-deterministically in the central cell, and sent out to the peripheral cells. We will show a method of communication between all the cells that allows the efficient simulation of bagging techniques for random forests.

References

1. L. Breiman, J. Friedman, R. Olshen, C. Stone. *Classification and regression trees*. Chapman & Hall. 1984.
2. D. Díaz-Pernil, F. Peña-Cantillana, M.A. Gutiérrez-Naranjo. Self-constructing Recognizer P Systems. In *Proceedings of the Thirteenth Brainstorming Week on Membrane Computing*, pp 137-154. Fénix Editora. 2014.
3. J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers. 1993.
4. C. Sammut, G. I. Webb, eds.: *Encyclopedia of Machine Learning*, Springer, 2011.
5. J.M. Sempere. A view of P systems from Information Theory. In *Proceedings of the 16th International Conference CMC16*. LNCS Vol. 10105, pp 352-362, Springer. 2017
6. J. Wang, J. Hu, H. Peng, M.J. Pérez-Jiménez, A. Riscos-Núñez. Decision Tree Models Induced by Membrane Systems. In *Romanian Journal of Information Science and Technology*, Vol.18, No. 3, pp 228-239. 2015
7. C. Zhang, Y. Ma, eds. *Ensemble Machine Learning*. Springer, 2012.