Pioneer Journal of Computer Science and Engineering Technology

Volume 6, Issues 1&2, Pages 1-18 (September & December 2013)

ISSN: 2231-184X

A COMBINATION OF GRAMMAR-BASED APPROACH AND MACHINE LEARNING FOR CLASSIFYING HERB SPECIES

T. Silwattananusarn, B. Martin, A. Mitrovic, J. Heinemann and W. Kanarkard

Received January 19, 2012

Abstract

DNA contains very long sequences that include repletion, noise and junk. This structure can affect the classification of the DNA sequence. In this paper, we propose the combination of grammar-based and machine learning approaches on a set of herb DNA sequences of Annoa squamosa (custard apple) and Zingiber officinale Roscoe (ginger). Five data sets with variable sequence length: 50, 100, 150, 200 and full length are used to evaluate and compare the performance of classifiers by their classification accuracy. The machine learning classification algorithms used are Decision Tree, Naïve Bayes, Support Vector Machine (SVM), and AdaBoost. The experimental results show that the best algorithms based on grammar-derived feature dataset are AdaBoost and SVM with an average accuracy of 85% and 75%, respectively.

Keywords and phrases: grammar-based approach, classification, classification algorithms.