Poster K-16

Predicting MEDLINE Abstracts Relevant to Salt Tolerance in Plants

Authors:
Crystal L. Nevels (*Department of Computer Science, Jackson State University, Jackson Mississippi, USA*)
Charles E. Bland (*Department of Computer Science, Jackson State University, Jackson Mississippi, USA*)
Raphael D. Isokpehi (*Department of Biology, Jackson Mississippi, USA*)

Short Abstract: We have built positive and negative corpora of MEDLINE abstracts of articles suitable for identifying words in abstracts that indicate the relevance of an abstract to salt tolerance in plants. A total of 952 positive predictive words were identified including salt, nacl, stress, tolerance, saline, osmoprotectant, drought, halophyte and mesembryanthemum.

Long Abstract:
High salinity in soils reduces plant yield and ultimately agricultural production. The ability of plants to grow in the presence of salinity is known as salt tolerance. High-throughput experiments such as sequencing of the genetic material (DNA) of organisms such as human, plants and microorganisms are yielding large quantities of data. The ultimate goal of conducting large-scale experiments is to translate these large amounts of information into useful knowledge that will find real world applications. However, an overwhelming increase in the literature discussing the experimental results and the rapid growth in content and number of biomedical literature databases present a major challenge of automatically identifying relevant information. Thus, the ability to automatically retrieve and classify this literature and extract pertinent information is a necessary step toward knowledge discovery. We have built positive and negative corpora of MEDLINE abstracts of articles suitable for identifying words in abstracts that indicate the relevance of an abstract to salt tolerance in plants. A corpus of 176 positive abstracts was built by extracting citations with MEDLINE links in the reference section of two review articles. Furthermore, a corpus of 3,379 negative abstracts were built based on the Gene Ontology annotation of Arabidopsis thaliana (3,271 articles) and Rice (108 articles). A word index containing all of the words in both the positive and negative articles was generated and the occurrence of the index term from the positive set and the negative set was determined. Using the word frequency index, the Chi-square test was performed to select features that specifically identified positive and negative articles. Furthermore, two measures positive predictive value (PPV) and the negative predictive value (NPV) were calculated to measure the percentage of articles that have positive features and the percentage of articles that have negative features respectively. A total of 952 positive predictive words were identified including salt, nacl, stress, tolerance, saline, osmotic, osmoprotectant, drought, halophyte and mesembryanthemum. These words will be incorporated in training sets for statistical and heuristic classifiers.