

Workshop: Data Mining in Agriculture 2010

Berlin, July 14th, 2010

Workshop Program

Starting 14:30 at the ICDM venue, 30 minutes for each presentation

Hierarchical Spatial Clustering for Management Zone Delineation

Georg Ruß, Martin Schneider, and Rudolf Kruse

An important task in classic agriculture is *base fertilization*. This term is commonly used to describe the process to make minerals like potassium (K), phosphor (P) and magnesium (Mg) available for the planted crops. Since the field is usually heterogeneous, the question arises which part of the field should be treated and how it should be treated. This is usually associated with the concept of *emphmanagement zone delineation*. There have been quite a number of approaches towards using fine-scale data for subdividing the field into a small number of zones. These approaches often require multi-year data sets and are based on low-resolution sampling methods for data

acquisition. Existing research into the semi-automatic generation of management zones usually applies a variety of clustering methods to subdivide the field into homogeneous parts. Most of the existing approaches use only parts of the available data for the clustering process. We propose a novel approach which tackles the above issues. We base our work on a site-year of high-resolution spatial data from selected wheat fields in Germany. Our approach may best be described as *emphdivide-and-conquer*: we split the field into homogeneous subfields and consecutively merge these subfields using a hierarchical agglomerative clustering approach with a spatial constraint.

Data Mining the Millennium Seedbank at Kew

Allan Tucker, Stephen Swift, Steve Counsel, Simon Kent, John Dickie, Kenwin Liu, and Robert Turner

Over 30,000 species of plant are edible, but we use only a tiny fraction of these in commercial agriculture. In the future we may well need a much greater range of species, particularly if climate change alters growing seasons or the world's population continues to increase and we run out of prime agricultural land. The Millennium Seed Bank (MSB) project is a ten year global conservation initiative, managed by the Seed Conservation Department of the Royal Botanic Gardens, Kew. This paper de-

scribes an initial application of data-mining techniques to this data, in order to generate rules or testable hypotheses and thus maximise its contribution to a germination decision support system with global applicability. We have shown that decision trees and Bayesian network classifiers can successfully illuminate the underlying relationships and direct further experiments due to the transparency in the way they partition the decision space.

Consistent Biclustering and Applications to Agriculture

Antonio Mucherino and Alejandra Urtubia

Consistent biclusterings of training sets can be exploited for solving classification problems in data mining. This technique has been mainly applied so far to solve classification problems related to gene expression data. However, it can be successfully applied to problems arising in other domains, and it is also able to provide information on the features causing the classification of the training set. We provide a quick overview of this technique

which is based on the concept of consistent biclustering, and we present a study on a particular problem arising in the agricultural field. We consider the problem of predicting problematic fermentations of wine at early stages of the vinification. The presented computational experiments show that the considered technique is able to provide some clues on the possible features causing the problematic fermentations.

Automated Vision-Based Diagnosis of Cassava Mosaic Disease

Jennifer Aduwo, Ernest Mwebaze, and John Quinn

Cassava Mosaic Disease (CMD) has been an increasing concern to all countries in sub-Saharan Africa that depend on cassava for both commercial and local consumption. Information about the country-wide spread of this disease is difficult to obtain, however, due to logistics and human resource issues in these countries. The objective of this study was to assess the applicability of an automated computer vision based diagnosis of CMD. Images of 92 healthy cassava leaves and 101 mosaic-

infected leaves were taken at Namulonge Crop Resources Research Institute, Uganda. We performed classification on these images based on shape and colour features in the images, and found that a Naive Bayes classifier operating on a combination of hue distribution, scale invariant feature transforms (SIFT) and speeded-up robust feature (SURF) transforms was able to perform near-perfect classification, with area under ROC curve of 99.56%, evaluated with 10-fold cross validation.