DATA MINING OF COTTON FIBER QUALITY MEASUREMENTS

Chris Turner
Hamed Sari-Sarraf
Texas Tech University, Electrical and Computer Engineering Department
Lubbock, TX
Eric F. Hequet
International Textile Center and Dept. Plant & Soil Science, Texas Tech University
Lubbock, TX

Abstract

The need for additional information (AFIS fiber length distributions, FTIR, etc.) in cotton breeding programs is leading to a new challenge: data warehousing and knowledge discovery in databases (KDD) also called data mining. Over the last few years laboratories at Texas Tech University International Textile Center and Cotton, Inc. have collected data on hundreds of thousands of samples from various cotton fiber quality breeding programs. Collectively this data largely remains an untapped resource of information. As such, we have begun to develop the tools necessary to analyze and process this data. In the past these cotton fiber quality measurements have been stored in a proprietary database program limiting accessibility. We, therefore, began by importing the data into a Microsoft SQL Server database, which is ODBC compliant and allows database connectivity from a wide range of applications and development platforms. With the information in a programmatically accessible location, we have developed a prototype data analysis tool using Matlab implementing a dynamic clustering method based on kernel density estimates and level set methods. This particular clustering technique has several advantages over some of the traditional unsupervised clustering methods including the ability to automatically determine the number of clusters present. Another interesting aspect of the algorithm is its treatment of outliers, as they tend to get clustered separately from the main densities. This could be particularly useful in identifying genetic mutants or simply maintaining quality control of the data.