Discovering hidden structures in high dimensional data space

ABSTRACT

The large volume of data that is currently collected in various fields of application can not be managed using data mining standard techniques.
The main purpose of this work of thesis is to find the most reasonable solutions for data mining problems related to the management of high dimensional data.

In particular two main applications of mining high dimensional data are considered in this work.

The first one deals with cloud detection, a problem of multispectral satellite image classification, demonstrating the high reliability of the statistical techniques of discriminant analysis in classifying this type of images. Such classification technique has been compared with standard ones
based on physical principles in order to benchmark the processing costs and the pass/fail rate.

The second application addresses the need to handle high dimensional data for which it is necessary to make assumptions rather than to have a confirmation (as in the previous application) . This naturally leads to the problem of clustering the data allowing to find significant structures within them. Instead of dwelling on one or more particular techniques of clustering, we chose to address the problem in a more comprehensive way by the so-called consensus clustering: rather than seek a single solution to the problem, the goal is to find all possible equivalently valid solutions. To this purpose an automatic procedure based on Least Squares Consensus Clustering has been developed.

The applications have been tested using both synthetic and real data-sets, actually demonstrating the validity of the procedures. Strong emphasis has also been put on results validation through the use of "goodness" indicators in order to demonstrate the reliability of the techniques developed.