Mining the FIRST astronomical survey

Imola K. Fodor and Chandrika Kamath
Lawrence Livermore National Laboratory

October 17, 2001

The Faint Images of the Radio Sky at Twenty-Centimeters (FIRST) is an on-going astronomical survey designed to cover over 10,000 square degrees of the North and South Galactic Caps. The data is collected with the NRAO Very Large Array in New Mexico, and is archived automatically into a searchable collection of images. At the 1 mJy source detection threshold, there are about 90 radio emitting sources per square degree, prohibiting exhaustive visual exploration of the data set.

In this presentation, we describe our results in applying semi-automated data mining techniques to detect galaxies with a bent-double morphology in the FIRST survey. Our main approach is to use supervised learning techniques, such as decision trees and linear models, to classify previously unseen galaxies based on models constructed from a labeled training set provided by the astronomers. We illustrate the entire data mining process, from explaining the data, through the extraction of features, to pattern recognition and interpretation of the results. We explain why defining and extracting meaningful features that discriminate bent-doubles from non bent-doubles is a non-trivial task, even after numerous consultations with our astronomer collaborators. Next, we detail our statistical and exploratory data analysis approaches to reduce the dimensionality of the feature set. After this step, irrelevant features and redundancies among the features are eliminated. Finally, we use the remaining features and the training set in conjunction with pattern recognition algorithms to build classifiers with the desired accuracy. The last stage is then to apply the best classifiers to predict the class bent, non bent of previously unclassified galaxies.

At the end of the data mining process, we provide the astronomers with a list of galaxies that are highly-likely to be bent-doubles. The astronomers can explore this smaller set first, instead of searching through the entire sky survey. The real power of data mining techniques is best illustrated by the fact that we found a bent-double that the astronomers missed in the manual search that generated the training set.

Our experience indicates that, if steered by appropriate domain knowledge, data mining methods can enhance the visual identification of bent-double galaxies. We found that careful extraction and selection of features, as well as feedback from the domain scientists, are essential for the successful application of data mining techniques.

UCRL-JC-145672 ABS. This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract W-7405-ENG.