The classification and regression tree project

January 16, 2019

- Describe the problem to be solved and your dataset. How was the raw data generated? What applications does the problem have? What predictors, that is attributes, are there and how are they converted to suitable input for your chosen tree learning algorithms?

  If you have many numerical attributes, consider using PCA for preprocessing. If you use PCA, you will need to scale and center each relevant attribute and reduce skewness using Box-Cox transformations.

  Also consider other ways to perform feature extraction and preprocessing. See Chapter 3 in the course book.

- What have others done previously with the same or similar data sets? Search for references on `scholar.google.com`.

- The easiest tree tools to use are C5.0 and Cubist. You should start your project using these.

  If you feel ambitious, you may then consider using random forest (rf) and gradient boosting (xgboost) accessed through the Caret package in R.

- Are trees or rules best? Try to explain the differences if there are any.

- Interpret the output from your tree tool(s) for the data set. Choose a number of rules and see which are sensible and which are not.

- Try to characterize the generalizing ability of the models generated by the tree software using repeated cross-validation or alternatively a test data set. How sensible to missing attributes or less training data is tree learning in your case?

- Do classifications have differing costs? If that is the case, use a `.costs` file in C5.0. Explain and analyze the result.

- Examine other relevant tree algorithm options for your data set, for example winnowing, boosting or pruning. Describe each option.

- Evaluate how good results you have obtained. It is more important to give a correct evaluation and work systematically than to obtain the lowest possible error percentage. What future improvements are there?