CSCD 429/529 Data Mining Homework #2 (40 Points)

Prediction of gene/protein localization

Data Set Description: This dataset was used in the 2001 kdd cup data mining competition. (http://www.cs.wisc.edu/~dpage/kddcup2001/). There were in fact two tasks in the competition with this dataset, the prediction of the "Function" attribute, and prediction of the "Localization" attribute. **Here we focus on the prediction of "Localization"** (this is somewhat easier as genes can have many functions, but only one localization, at least in this dataset). The dataset provides a variety of details about the several genes of one particular type of organism. The main dataset (*Genes_relation.data* and *Genes_related.test*) contains row data of the following form:

Gene ID, Essential, Class, Complex, Phenotype, Motif, Chromosome Number, Function, ***Localization*****.**

The description of data attributes was given in file *Genes_relation.names*. The first attribute is a discrete variable corresponding to the gene (there are 1243 gene values). Also the remaining 8 attributes consist of discrete variables, most of them related to the proteins coded by the gene, e.g. the "Function" attribute describes some crucial functions the respective protein is involved in, and the "Localization" is simply the part of the cell where the protein is localized.

In addition to the above files, there are also data files (*Interactions_relations.data* and *Interactions_relation.test*) which contain information about interactions between pairs of genes.

Data File Size:

- Gene_relation files: 6275 examples (4346 training, 1929 test), 8 categorical attributes.
- Interaction_relation files: 1806 records, 2 attributes (one categorical; one numerical)

Task: **Build a classification model to predict attribute "Localization".** Detailed knowledge of the biology should not be necessary for this assignment. One word of caution: **your classifier for localization should not use "function"**, since **both** fields will be withheld from the test genes when they are provided.

Challenge: This dataset is a great challenge. One issue is that there is a high proportion of missing variables in the Genes_relation data. The other issue is how to use the interaction data effectively.

Keys: The keys are provided in the file keys.txt. Use this file to evaluate the accuracy of your solution.

References:

- Talk overview slides about this problem and also the winner presentation in the KDD 2001 competition can be found on-line.
- See also Answers to Questions of General Interest from Question Period 1 and Answers to Questions of General Interest from Question Period 2

Deliverables:

- (35 points) All workable program files: in this assignment, you are required to build a classification learning model using one of the approaches we covered in **THIS** class to predict gene localization. You can choose from the following list of classification algorithms:
    - Decision Tree
    - Naïve Baye
    - K-nearest neighbor

  Note 1 to both CSCD 429 and 529 students: you may implement the algorithm using any programming language of your choice. **However, you must implement the underlying classification algorithm from scratch.** In other words, you are NOT allowed to use existing machine learning related libraries in your program to build the learning model. You need to clearly document all the references you have used in this assignment.

  Note 2 to CSCD 529 students: you need to develop a solution that integrates the ***interaction*** file as the input data.

- (1 point) A result (i.e. predition) file in the format of **<gene ID>, <localization>** in each row.
- (4 points) A report that includes
    - a list of references you used in this assignment;
    - a description on how to run your program;
    - a brief description on the classification method you used and how you handled the missing data and/or interaction data;
    - the accuracy you obtained and a description on how you calculated the accuracy. **The overall accuracy generated by your solution is expected to be higher than 50%.**
- Include all the files into a single .zip file and **submit your file via Canvas**.