<Sommelier>

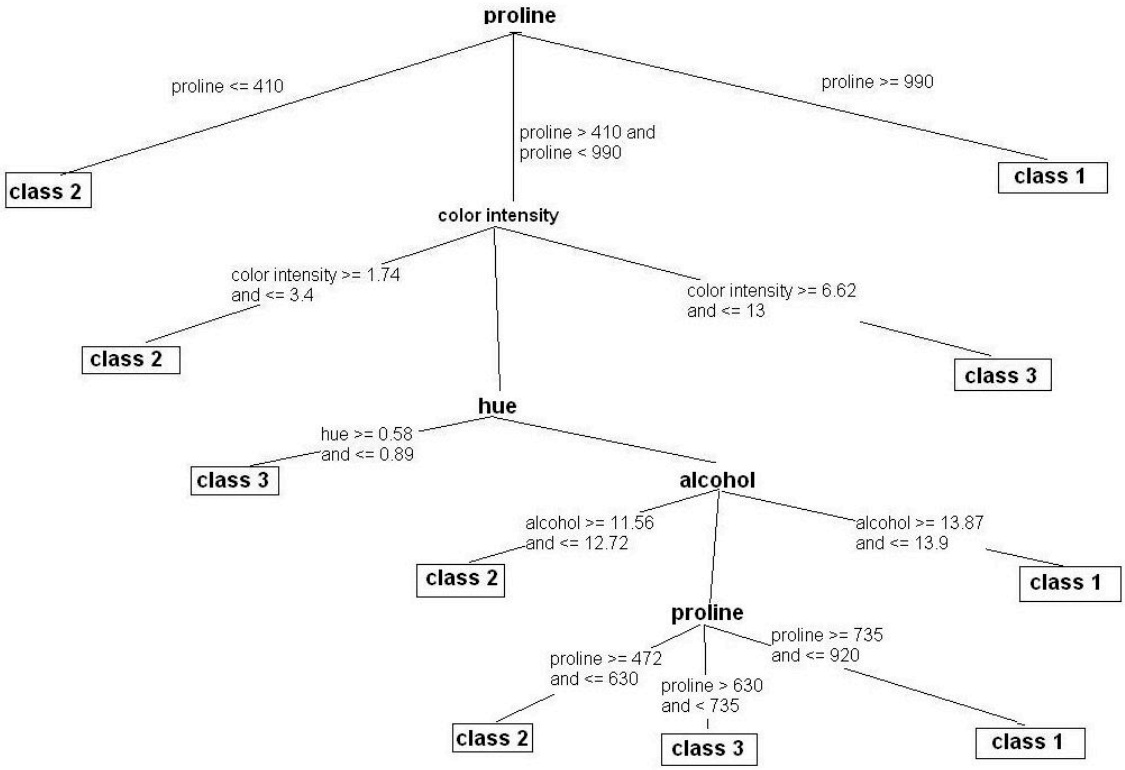
(Khoa Doan)

Executive Summary

An automated chemical analysis device is used to classify wine types. In order to do this, it must have a classification model. This model is created on a training data set containing 153 rows. Each row, in turn, has 13 values of the wine properties, which are alcohol, malic acid, ash, alkalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols, proanthocyanins, color intensity, hue, OD280/OD315 of diluted wines and praline. The dataset also has a wine type associated with each row.

The approach that is chosen to partition the dataset is called decision tree. It begins with determining an attribute or wine property which provides the most useful information to dissect the dataset. Praline is chosen here since it can partition the set into more useful subset than any other attributes. In fact, it divides the dataset into three subsets: two can be clearly classified as class 1 and 2, and one contained conflicted records. This unsolved subset is, then, use the partitioning process again until all of its child subsets can be clearly classified.

Applying the decision tree approach to the dataset, a model is created below for classifying wines. The model is fairly complete since it provides 100% classification accuracy if the dataset is assumed to correctly provide the classification model.



This model can be used by the automatic chemical analysis machine to classify wines. The bolded properties above are the values input at each stage in the classification process. The condition at each branch indicates the criteria of the branching to either a wine class or a deeper level.

Problem Description

There are three types of wine, labeled as 1, 2 and 3. Each type has eighteen properties, which are alcohol, malic acid, ash, alkalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols, proanthocyanins, color intensity, hue, OD280/OD315 of diluted wines and praline.

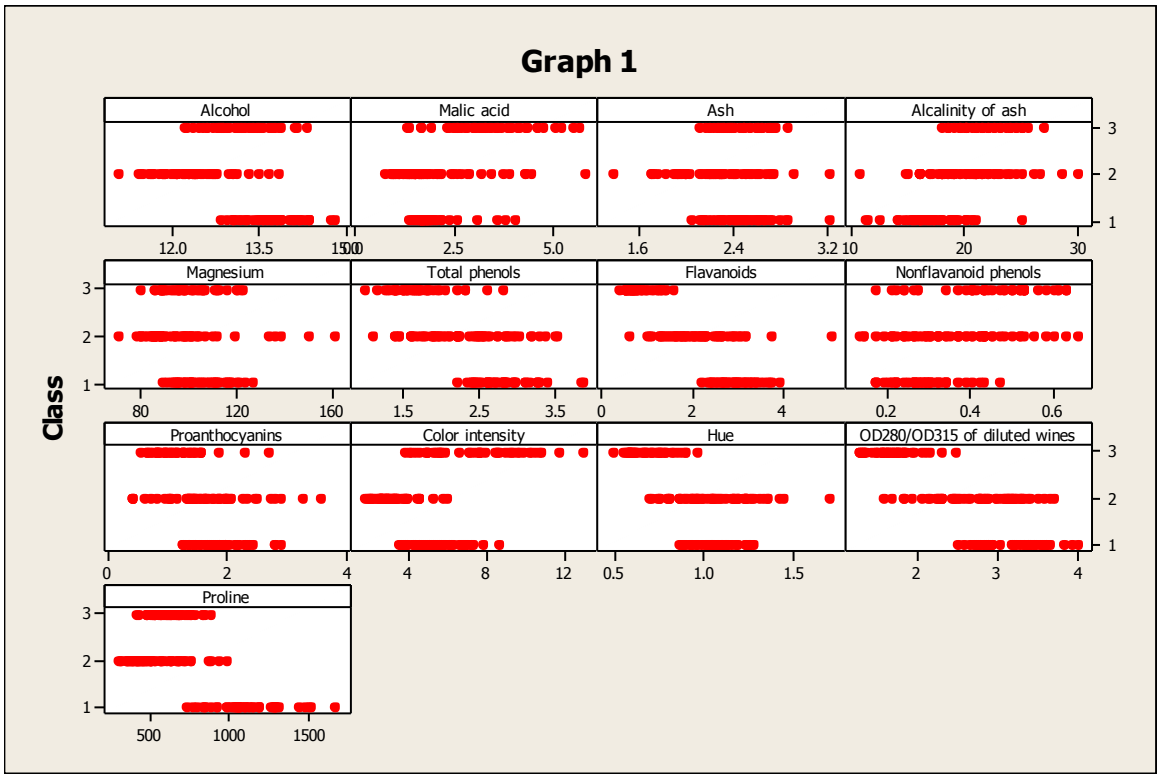
Revised: 13 December 2003

These properties, as well as types of wine, are included in a training dataset which has a total of 153 records. This dataset is descriptively analyzed to recognize an attribute or set of attributes that can be used by an automatic chemical analysis device to classify wines.

Analysis Technique

The classifying process uses the decision tree approach to divide the source set into subsets, each of which will map to a wine class. The approach starts with identifying the splitting attributes or the input attributes around which the divisions will take place. If a subset clearly identifies a class, the splitting stops. Otherwise, the process is repeatedly applied on the subset until all of its child subsets are clearly classified.

The most important task in decision tree approach is to determine the splitting attributes at each step. Here, each attribute of the wine dataset is graphed against the wine type in order to identify the splitting attributes that provide the most information. In order word, the graph with the least overlapping regions will be more useful to separate the dataset.



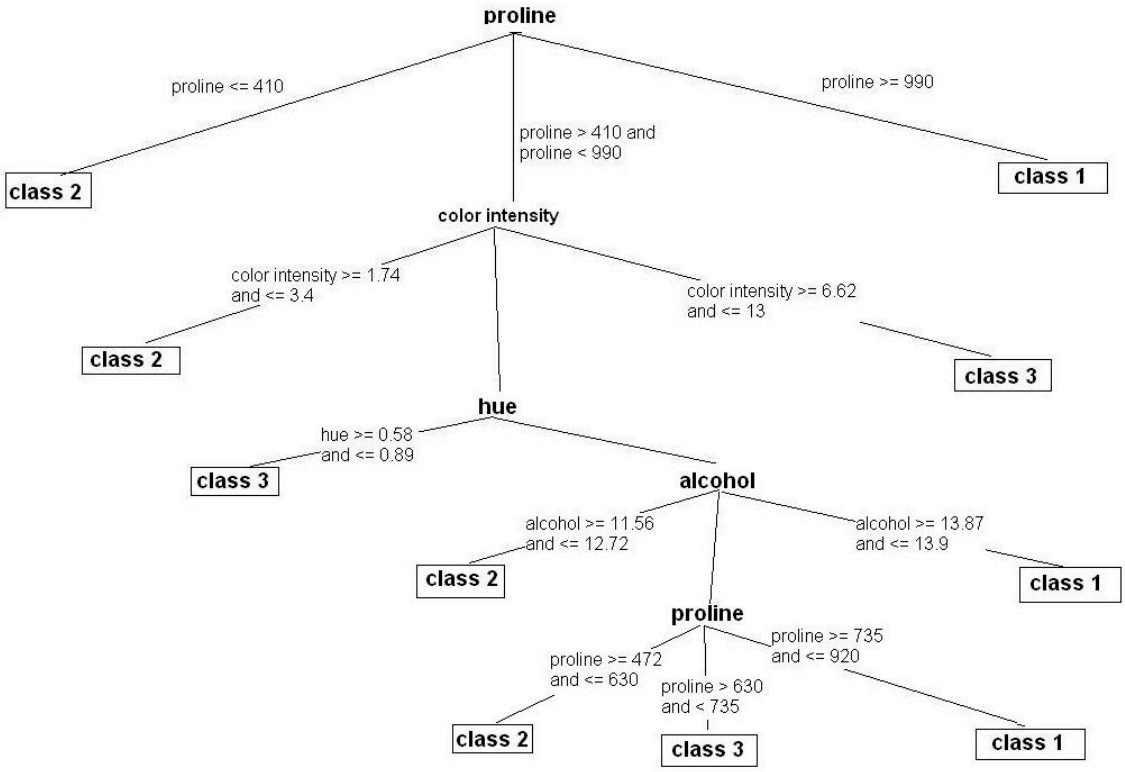
Graph 1 suggests that praline can be used as the splitting attribute. Then, the dataset is sorted on praline and is divided into three parts. Two of them are at the top and bottom (both yellow in the tables below) of the set, when the wine classes just change from one to another; and they are mapped to wine class 2 and 1. In addition, if the at the boundary, the values of the splitting attribute in the 2 parts are the same, the new boundary is created at the above or below values that are next largest or smallest accordingly as in the tables below.

13.9	1.51	2.67	25	86	2.95	2.86	0.21	1.87	3.38	1.36	3.2	410	2
11.8	1.72	1.88	20	86	2.5	1.64	0.37	1.42	2.06	0.94	2.4	415	2
13.9	5.04	2.23	20	80	0.98	0.34	0.4	0.68	4.9	0.58	1.3	415	3

13	1.67	2.6	30	139	3.3	2.89	0.21	1.96	3.35	1.31	3.5	985	2
13.9	1.68	2.12	16	101	3.1	3.39	0.21	2.14	6.1	0.91	3.3	985	1
13.7	1.83	2.36	17	104	2.42	2.69	0.42	1.97	3.84	1.23	2.9	990	1

Revised: 13 December 2003

The interpretation of the first partitioning is: if a record has a praline value less than 470, it can be classified as class 2; on the other hand, if a record has a praline value greater than 990, it can be mapped to class 1; and the other subset contains unresolved classification. The iterative process is again applied for this unresolved subset until all of its children subsets are mapped to a particular wine class. A complete classification tree is below:



The accuracy of this classification is 100% since all records can be mapped to a wine class successfully. This, in turn, can be used as a classification model for the chemical analysis device to classify wine types based on praline, color intensity, hue, and alcohol attribute values.

Assumptions

Here are some assumptions on the model:

- The training dataset can be used as a representative for classifying wines.
- Input conditions are fairly correct to provide a useful model.

Results

The decision tree approach produce a model, as above, that can be used to classify wine types. Here, the node attributes that are not at the leaf of the tree are the input at each stage. The tree's arcs are labeled with conditions for classification. The arcs without condition represent all other possible values that are not in sibling arc's conditions. Based on this model, the automated chemical analysis device can provide a classification function to the sommelier.

Issues

The chosen splitting-attributes set are not the only one in decision tree approach. Choosing the right splitting attributes is not an easy task. Here, they are chosen based on useful information they provide, and this has created a fairly useful classification model. In addition, by dividing the dataset into region, we can also resolve the fact that decision tree cannot handle continuous data.