Plant identification by leaf shape

Wei Liu

There are estimated to be nearly half a million species of plant in the world. Considering such huge number of species, sometimes, identifying and classifying different plants is a challenging work even for the plant specialists. Among the characteristics of the plants, for example, size, leaf shape, flower color, odor, form, etc., leaf shape is considered a good one for the classification and data mining. In my proposal, based on what I've learnt so far, I will build different classifier for this plants species classification/identification problem. Also, the classifiers will be applied on the test dataset, and the performance of different classifiers will be evaluated.

First, the training dataset, which is leaf shapes and corresponding tree species are provided. Key features can be extracted from each leaf shape and presented by row of data (the feature extraction from image is also an interesting topic in data mining and there are lot of resources for learning online). The training data and test data I will use can be obtained from kaggle. There are in total 192 attributes for each instance. In the training data, there are 1584 instances.

The algorithms I plan to use are 1) decision tree and 2) naïve Bayes. The evaluation criteria are so called logloss as follows:

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} log p_{ij}$$

N is the number of instances in the test set, M is the number of attributes, y_{ij} is 1 if observation i is in classes and 0 otherwise, p_{ij} is the predicted probability that observation i belongs to class j.

My proposed work includes the following:

1) Use the packages (e.g., scikit-learn) of python to test the classifiers;

2) Write my own python code to do the same task, thus in this way to deepen understanding to the algorithms, also to practice my coding skills.

3) Use the same classifiers in weka to do the same analysis and evaluate its performance.

In class, we have already learnt the principals and mechanisms for different data mining algorithms. Based on those knowledge, from this project, I wish I can get the training in coding, especially get myself familiar with data mining/machine learning packages in python and also the data mining tools with weka.