Sound Creation Lab

Lab 3: Sound Classification

Goal

The goal of this lab is to learn and practice the automatic classification of sounds using the Weka data mining software.

Evaluation

This lab will be evaluated based on the completion of the practical tasks (1–11) and the submission of a written report. The report must be in PDF format and include your name(s), student ID(s) and the title of the lab. The report must be submitted via Moodle by the date indicated on the calendar.

Task 1

Download the sounds for this lab from Moodle: `Classification_sounds.zip`
The files are divided into three classes: listen to a few files from each class and provide a description of each of the three classes of sounds and their perceptual characteristics. [0.5 point]

Task 2

Download and install the Weka data mining software (make sure it's not already installed before proceeding): `http://www.cs.waikato.ac.nz/ml/weka/`
Download the descriptor file `descriptors.csv` from Moodle. this file contains the summary statistics (max, min, mean, median and standard deviation) for all descriptors available in the MIR.EDU vamp library (with the exception of the the attack start/end times).

Task 3

Open the Weka Explorer, go to the 'Preprocess' tab and open the descriptor file.

1. How many instances are there in total? [0.25 point]
2. How many attributes are there? [0.25 point]
3. How many classes are there? [0.25 point]
4. How many instances are there in each class? [0.25 point]

Task 4

Remove attribute no. 1 (filename). Next, normalize the values of each feature by applying the following filter: `weka > filters > unsupervised > attribute > Normalize`. Normalizing your features can improve the classification performance of some classifiers. Now go to the 'Vizualize' tab in Weka. Examine the plots for the descriptor pairs mentioned below. For each pair display a screenshot of the plot and comment on the distribution of the instances:

1. X: Temporal Centroid, Y: Spectral Spread std [0.5 point]
2. X: Spectral Flux std, Y: Spectral Skewness Max [0.5 point]

3. X: Spectral Centroid mean, Y: Zero Crossing Rate Mean [0.5 point]

Task 5

What is cross validation? When might we want to use cross validation? [0.5 point]
(`http://en.wikipedia.org/wiki/Cross-validation`)

Task 6

Go to the 'Classify' tab. Now classify the data using the ZeroR (baseline) classifier and 10-fold cross validation (test option). What classification accuracy do you get? Display the confusion matrix in your report and explain it. [0.5 point]

Task 7

Repeat the classification using the J48 (decision tree), SMO (support vector machine) and Ibk (k-Nearest Neighbours) classifiers. For each classifier:

- Report the classification accuracy.
- Display the confusion matrix and explain it. For J48 also display the classification tree.
- Identify between which two classes there is the most confusion.

Which classifier provides the highest classification accuracy?
[3 points]
(`en.wikipedia.org/wiki/Decision_tree`)
(`en.wikipedia.org/wiki/Support_vector_machine`)
(`en.wikipedia.org/wiki/K-nearest_neighbor_algorithm`)

Task 8

Change the minimum number of instances per leaf (`minNumObj`) in the J48 classifier from 2 to 10 and repeat the classification:

- How does this affect the classification tree? [0.25 point]
- How does this affect the classification accuracy and why? [0.25 point]

Task 9

What is feature selection? What is the motivation for applying feature selection? [0.5 point]
(`en.wikipedia.org/wiki/Feature_selection`)

Task 10

Go to the 'Select attributes' tab in Weka. Perform feature selection using the `CfsSubsetEval` attribute evaluator and the `BestFirst` search method with the default parameters. Use the full training set.

- How many features did you have in total prior to selection?
- How many features were selected?
- Which features were selected?

[0.5 point]

Task 11

Go back to the 'Preprocess' tab and remove all features except for those selected by the feature selection process and the class label.

- Repeat the steps of Task 7 using the selected features. [1 point]
- Comment on the difference in results when using all features and when using only the selected features. [0.5 point]