Feature selection using mutual information software

Fernando de la Calle Silos

1 Introduction

We have developed an automatic method for feature selection that optimizes the subset of features used on a classification problem.The method is based on the measure of the Mutual Information (MI) between the features and a ground truth vector. Obviously, due to the need of ground truth labels, this procedure can only be applied in supervised scenarios.

The mutual information $I(X;Y)$ between two random variables X and Y measures the mutual dependence between them; in other words is the amount of uncertainty about Y that is removed by knowing X: $I(X;Y) = H(X) - H(X|Y)$ where H is the entropy of a variable. The mutual information was calculated using MILCA algorithm [1]. Furthermore, details about the implemented method are given in the Algorithm 1.

Algorithm 1 Feature selection using mutual information.

1: Start with the empty set X^0 and consider an initial value of mutual information $MI^0 = 0$.
2: **for** $t = 1 \rightarrow N_f =$ Number of features **do**
3: **for** each feature i not included in the set **do**
4: Compute MI between the extended set $X_i^t = \{X^{t-1}, X_i\}$ and the ground truth vector: $I(X_i^t; Y)$.
5: Compute the increment on the MI as $\Delta MI_i = MI_i^t - MI^{t-1}$.
6: **end for**
7: **if** $\Delta MI_i == 0$ for every i **then**
8: The feature set X_{t-1} is selected and the algorithm ends.
9: **else**
10: Select the feature $X*$ that maximizes ΔMI_i and add it to the set $X^t = [X^{t-1}, X^*]$
11: **end if**
12: **end for**

2 Instructions

The MATLAB program runs calling the following function:

```
[IM, selecFeatures] = IMFeatureSelec(features, label);
```

where `features` are a $1 \times M$ cell, containing the M features to analyze, and `label` are a $1 \times N$ ground truth vector. Each feature contained on the `features` cell must have dimension of $k \times N$.

The exit of the program are a IM matrix containing the computed mutual information for each step and a *selecFeatures* vector with the selected features. Also the program printed all this data on the command line:

```
-------Feature selection using mutual information-------
                Fernando de la Calle Silos
             Universidad Carlos III de Madrid
```

```
-------------------------------------------------------

The selected features are:
Feature 1
Feature 2
Feature 3

   Mutual Information Table
-------------------------------------------------------
Feature 1 0.0925   ---    ---    ---    ---
Feature 2 0.0859 0.1338  ---     ---    ---
Feature 3 0.0525 0.0813 0.1398   ---    ---
Feature 4 0.0456 0.0821 0.1322 0.1329   ---
Feature 5 0.0867 0.1006 0.1051 0.0980 0.1012
```

A test script *test.m* and a data file *data.mat* are included to test the algorithm.

The mutual information was calculated using the MILCA (*Mutual Information Independent Component Analysis*) algorithm [1]. This can be download from: `http://www.klab.caltech.edu/~kraskov/MILCA/`, and must be added to the directory called: `MutualInformationICA`.

You maybe have to compile the C-source codes into executable files, which will be then called be the MATLAB programs. To do this under linux or MacOS you have to type in the terminal the following commands:

```
icc -c miutils.C -o miutils.o
icc  miica.C -o milca miutils.o
```

The script `Tutorial.m` inside the folder `MutualInformationICA` also explains this, and includes some examples about computation of mutual information.

References

[1] Harald Stögbauer, Alexander Kraskov, Sergey A. Astakhov, and Peter Grassberger. Least-dependent-component analysis based on mutual information. *Phys. Rev. E*, 70:066123, Dec 2004.