aVersion 0.4
Updated 18.2.2014

Description of weakly supervised learning dataset (WSLD) and the associated learning challenge

Okko Räsänen[1], Unto K. Laine[1] and Jukka P. Saarinen[2]

[1]*Department of Signal Processing and Acoustics, Aalto University, Finland*
[2]*Nokia Research Center (NRC) Tampere, Finland*

1. Introduction

This document describes a compact dataset (weakly supervised learning dataset, or WSLD) that can be used to evaluate and compare weakly supervised learning algorithms performing on sequential categorical data. The goal is to facilitate methodological development in the field and to enable benchmarking of different algorithms

Weakly supervised learning refers to a machine learning paradigm where annotation of the data is not precise. Instead, each training signal $\mathbf{s} \in S$ of training set S is associated with a set of class labels $\mathbf{c} = \{c_1,c_2,\ldots,c_N\}$, $c \in C$, that denotes the possible presence of patterns $\mathbf{c}$ in the signal, but without the accurate knowledge of locations or temporal order of the patterns. The goal of a weakly supervised learning method is to learn classifiers for the patterns C in the data, allowing the recognition of the patterns from new input without the label information.

Weakly supervised learning is typical to many real-world learning situations where associations between two or more data streams have to be learned without explicit guidance. As an example, the word learning process of a human child can be modeled as a weakly supervised cross-situational learning problem (see, e.g., Räsänen & Laine, 2012) where visual objects (labels C) have to be associated to their correct acoustic word forms (data S). In similar manner, weak labeling can originate from processes where, e.g., medical or process industry data are known to be associated to some external explanatory variables, but the relationship between these two modalities are not precisely understood.

The next section describes the data set associated with this document (also available from http://www.acoustics.hut.fi/~orasanen/WSLD) and section 3 describes the experimental setup that can be used to evaluate algorithm performance using the data set. Section 4 shows the baseline results in the learning tasks, and section 5 contains information regarding submitting your own results to be included in the future versions of the document.

2. Data

The WSLD dataset consists of 3337 data sequences of varying length. Each sequence is associated with 4 labels (unordered, unaligned) denoting the presence of a corresponding pattern in the sequence. There are a total of 50 unique patterns ($|C| = 50$). The patterns of interest are known to be distributed in time, spanning multiple sequence elements. In addition, it is known

that the patterns of interest are interleaved in the data. This means that the training and evaluation sequences contain extraneous content (noise) irrelevant to the classification problem.

The data is divided into a training set of 2500 signals and an evaluation set of 837 signals.

The sequential data comes in two resolutions (two separate tasks): a low-resolution set (LRS) with vocabulary size of 12 unique elements, and a high-resolution set (HRS) with 128 unique elements.

Data is provided in both MATLAB (.mat) format and in ASCII (.txt) format. For MATLAB, there are two files, LRS.mat and HRS.mat, corresponding to the low-resolution set and high-resolution set, respectively. Both .mat files contain the following variables:

data_train	(cell array)	training sequences
data_test	(cell array)	evaluation sequences
labels_train	(2500x4 matrix)	training labels
labels_test	(837x4 matrix)	evaluation labels

The ASCII files are organized into two .zip files, LRS.zip and HRS.zip. Both .zip files contain four .txt files organized according to the variables above, filename denoting the variable stored in the corresponding file. Each row corresponds to one signal (sequence elements separated by whitespace, signals separated by new line).

3. The experimental setup of the challenge

In the challenge, the goal is to use the training set of 2500 signals to train classifiers for the correspondences between signal patterns and the labeling. Only the training set sequences and the associated labels can be used in the training stage.

During evaluation, the classifier is asked to provide 4 pattern hypotheses $\mathbf{c} = \{c_1,c_2,c_3,c_4\}$ for each test sequence. These 4 hypotheses are then compared to the underlying ground truth $\mathbf{c}_{\text{true}}$ with the same number of labels. All hypothesized patterns also included in the ground truth are considered as correctly recognized. The recognition rate is defined as the number of correctly hypothesized pattern labels divided by the total number of test patterns.

$$\boldsymbol{N}_{\text{correct}}/N_{\text{total}} = N_{\text{correct}} / 3348. \qquad (1)$$

Since the number of patterns in the test data is always known, this a simplified *classification*–based evaluation procedure in comparison to the task of *pattern detection*. However, it allows easier and faster benchmarking of algorithms without the necessity to optimize performance in terms of detection threshold.

The challenge is divided into two sub-challenges according to the division of the dataset: high-resolution and low-resolution challenges using the LRS and HRS data sets, respectively. Performance in both sub-challenges is evaluated separately. Optimization of algorithm parameters is allowed between the sub-challenges.

4. Baseline results

The baseline results were computed using the Concept Matrix (CM) algorithm of Räsänen & Laine (2012) and Non-linear Mapping CM (NMCM; unpublished experimental version)

CM algorithm was operated using lags **k** = {1, 2, …, 15}, maximum lag estimated according to the shape of the mutual information function (see Räsänen & Laine, 2012 for details). Results for both algorithms are shown in Table 1.

Table 1: Baseline results for the WSLD dataset.

	CM	**NMCM**
Dataset	% correct	% correct
LRS	37.87 (± 0.00 %)	**55.59** (± 0.00 %)
HRS	79.51 (± 0.00 %)	**79.69** (± 0.00 %)

Algorithm running times (MATLAB 2011a in OS X, 4x 3.2 GHz Intel Xeon, 12 GB RAM; training only) are reported in Table 2. Note that the reported times are only suggestive: none of the algorithms were strictly optimized for speed and computational performance on different platforms may vary.

Table 2: Running times of the algorithms in the training stage.

	CM	**NMCM**
Dataset	time (s)	time (s)
LRS	54.7	236.4
HRS	59.9	256.0

5. Reporting results

It is desirable to accumulate understanding of how different learning algorithms perform in the provided weakly supervised learning tasks. If you wish to report your results to be included in the future versions of this document, please contact Okko Räsänen (okko.rasanen@aalto.fi). In addition to the results in either or both of the sub-challenges, please provide a brief description of the methodology used in the experiments including possible deviations from the standard versions of well-known algorithms. Your results can be reported with or without author name upon request.

6. References

Räsänen O. & Laine U.: "A method for noise-robust context-aware pattern discovery and recognition from categorical sequences", *Pattern Recognition*, Vol. 45, pp. 606-616, 2012

Räsänen O., Leppänen J., Laine U. & Saarinen, J.: "Comparison of Classifiers in Audio and Acceleration Based Context Classification in Mobile Phones", *Proc. EUSIPCO'11*, Barcelona, Spain, pp. 946-950, 2011

Version information

v0.4	18.2.2014	Removed NMF baseline results as it is likely that better NMF results can be obtained with a proper state-of-the-art implementation (please contribute).
v0.3	15.5.2012	Added NMF baseline results and computation time estimates.
v0.2	7.5.2012	Dataset published