A Toolbox for Adaptive Sequence Dissimilarity Measures for Intelligent Tutoring Systems

Benjamin Paassen
CITEC Center of Excellence
Bielefeld, Germany
bpaassen@techfak.uni-bielefeld.de

Bassam Mokbel
CITEC Center of Excellence
Bielefeld, Germany
bmokbel@techfak.uni-bielefeld.de

Barbara Hammer
CITEC Center of Excellence
Bielefeld, Germany
bhammer@techfak.uni-bielefeld.de

ABSTRACT

We present the *TCS Alignment Toolbox*, which offers a flexible framework to calculate and visualize (dis)similarities between sequences in the context of educational data mining and intelligent tutoring systems. The toolbox offers a variety of alignment algorithms, allows for complex input sequences comprised of multi-dimensional elements, and is adjustable via rich parameterization options, including mechanisms for an automatic adaptation based on given data. Our demo shows an example in which the alignment measure is adapted to distinguish students' Java programs w.r.t. different solution strategies, via a machine learning technique.

1. INTRODUCTION

Systems for computer-aided education and *educational data mining* (EDM) often process complex structured information, such as learner solutions or student behavior patterns for a given learning task. In order to abstract from raw input information, the given data is frequently represented in form of sequences, such as (multi-dimensional) symbolic strings, or sequences of numeric vectors. These sequences may represent single solutions, as in some *intelligent tutoring systems* (ITSs) [2, 6]; or may encode time-dependent data, like learner development or activity paths [1, 7].

Once a meaningful sequence representation is established, there are many possibilities to process sequential data with existing machine learning or data mining tools. A crucial component for this purpose is a (dis)similarity measure for pairs of sequences, which enables operations like finding closest matches in a given data set, clustering all instances, or visualizing their neighborhood structure [5]. One particularly flexible approach to determine the (dis)similarity of sequences is *sequence alignment* [3].

For applications in the context of EDM and ITSs, sequence alignment offers two key features: On the one hand, the structural characteristics of sequences are taken into account, while calculation remains efficient, even with complex parameterization options. On the other hand, alignment provides an intuitive matching scheme for a given sequence pair, since both sequences are extended, so that similar parts are *aligned*. However, we believe the full potential of sequence alignment is rarely utilized in EDM or ITSs.

Acknowledgments: Funding by the DFG under grant numbers HA 2719/6-1 and HA 2719/6-2 and the CITEC center of excellence is gratefully acknowledged.

2. ALIGNMENT TOOLBOX

We present the *TCS Alignment Toolbox*[1], an open-source, Matlab-compatible Java library, which provides a flexible framework for sequence alignments, as follows:

Multi-dimensional input sequences are possible, such that every element of the sequence can contain multiple values of different types (namely discrete symbols, vectors or strings).

A **variety of alignment variants** is implemented, covering common cases, such as *edit distance*, *dynamic time warping* and *affine sequence alignment* [3].

The **parameterization** of the alignment measure is defined by costs of operations (replacement, insertion, and deletion) between sequence elements, which can be adjusted by the user, or left at reasonable defaults. Users can even plug in custom functions to yield meaningful problem-specific costs.

A **visualization feature** displays the aligned sequences in a comprehensive HTML view, as well as the dissimilarity matrix for an entire set of input sequences.

An approximate **differential of the alignment functions** w.r.t. its parameters is provided, which enables users to automatically tune the rich parameter set with gradient-based machine learning methods, e.g. to facilitate a classification [4].

In this demo, we present an example for a set of real student solutions for a Java programming task: After programs are transformed to sequences, the parameters of an alignment algorithm are automatically adapted to distinguish between different underlying solution strategies, and the resulting alignments are visualized. Thus, the adapted measure improves the classification accuracy for the given data.

3. REFERENCES

[1] S. Bryfczynski, R. P. Pargas, M. M. Cooper, M. Klymkowsky, and B. C. Dean. Teaching data structures with besocratic. In *ITiCSE 2013*, pages 105–110. ACM, 2013.

[2] S. Gross, B. Mokbel, B. Hammer, and N. Pinkwart. How to select an example? A comparison of selection strategies in example-based learning. In *ITS 2014*, pages 340–347, 2014.

[3] D. Gusfield. *Algorithms on Strings, Trees, and Sequences.* Cambridge University Press, New York, NY, USA, 1997.

[4] B. Mokbel, B. Paassen, F.-M. Schleif, and B. Hammer. Metric learning for sequences in relational LVQ. *Neurocomputing*, 2015. (accepted/in press).

[5] E. Pekalska and B. Duin. *The Dissimilarity Representation for Pattern Recognition*. World Scientific, 2005.

[6] E. R. Sykes and F. Franek. A prototype for an intelligent tutoring system for students learning to program in Java (TM). *IASTED 2003*, pages 78–83, 2003.

[7] N. van Labeke, G. D. Magoulas, and A. Poulovassilis. Searching for "people like me" in a lifelong learning system. In *EC-TEL 2009*, volume 5794 of *LNCS*, pages 106–111. Springer, 2009.

[1] Available at `http://opensource.cit-ec.de/projects/tcs`