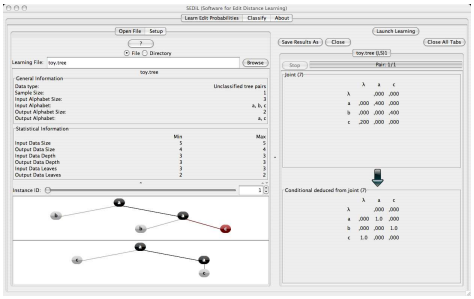
SEDiL: Software for Edit Distance Learning

http://labh-curien.univ-st-etienne.fr/SEDiL

Laurent Boyer(1), Yann Esposito(1), Amaury Habrard(2), Jose Oncina(3) and Marc Sebban(1)

(1) Laboratoire Hubert Curien, UMR CNRS 5516, Université de Saint-Étienne, France (2) Laboratoire d'Informatique Fondamentale, UMR CNRS 6166, Aix-Marseille Université, France

(3) Departamento de Languajes y Sistemas Informatico, Universidad de Alicante, Spain

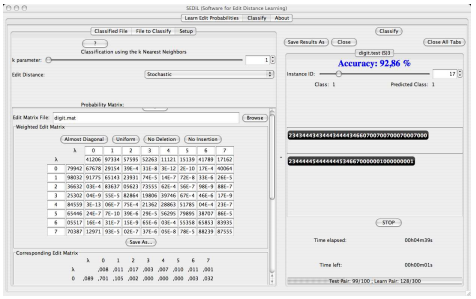


Standard Edit Distance

Context

- Structured data (Strings, Trees, Graphs)
- Recurrent need of similarity measures in Machine Learning
  - ↳ Edit Distance (ED)
- Theoretical framework
  - ↳ Probabilistic models (Pair-HMMs, transducers, PDFAs, ...)
- Aim
  - ↳ To provide a platform grouping together the state of the art algorithms for learning edit parameters

⇒ **SEDiL**



Probabilistic Edit Similarity

Definition

An **Edit script** is a sequence of edit operations (insertion, deletion, substitution) allowing the transformation of an input data X into an output data Y.

Definition

The **Standard Edit Distance (ED)** between X and Y is the cost of the less costly edit script.

Crucial step

Manual setting of the edit costs (depends on the domain).

Definition

The **Probabilistic Edit Similarity (ES)** between X and Y is defined as: ES(X,Y)=-log p(X,Y), where p(X,Y) is the (joint or conditional) probability of transforming X into Y, according to either the best script (Viterbi) or all the possible scripts (stochastic).

Crucial step

Learning of a distribution over the edit operations (automatic capture of the domain knowledge).

Example

pair of strings

input string a b
output string a a

Edit scripts = { (a → a , b → a) ; (a → λ , b → a , λ → a) ; (a → a , b → λ , λ → a) ; ... } where λ is the empty symbol

A priori fixed edit costs

c	λ	a	b
λ	-	2	3
a	3	0	1
b	1	2	0

ED(a b , a a) = c(a → a) + c(b → a)
= 0 + 2 = 2

Learned edit probabilities

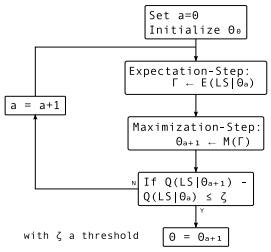
δ	λ	a	b
λ	—	0.2	0.1
a	0.05	0.0	0.1
b	0.05	0.2	0.3

p(a b , a a) = δ(a → λ) x δ(b → a) x δ(λ → a)
= 0.05 x 0.2 x 0.2 = 0.002

ES_λ(a b , a a) = -log(0.002)
≈ 2.69

ES learning algorithms in SEDiL

Expectation-Maximization-based algorithms



References

[1] Ristad S., Yianilos P.: Learning string-edit distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(5), (1998), 522-532.

[2] Durbin R., Eddy S., Krogh A., Mitchison G.: Biological sequence analysis. Cambridge University Press, (1998).

[3] Oncina J., Sebban M.: Learning stochastic edit distance: application in handwritten character recognition. Pattern Recognition, 39(9), (2006), 1575-1587.

[4] Bernard M., Habrard A., Sebban M.: Learning stochastic tree edit distance. ECML'06, (2006), 42-52.

[5] Boyer L., Habrard A., Sebban M.: Learning metrics beetween tree structured data: Application to image recognition. ECML'07, (2007), 54-66.

[6] Bernard M., Boyer L., Habrard A., Sebban M.:Learning probabilistic models of tree edit distance. Pattern Recognition, 41(8), (2008), 2611-2629.

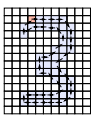
Features of the state of the art algorithms

[1]	[2]	[3]	[4,6]	[5]
strings	strings	strings	trees	trees
memoryless	non-memoryless	memoryless	memoryless	memoryless
Viterbi stochastic	stochastic	Viterbi stochastic	Viterbi stochastic	stochastic
generative	generative	discriminative	generative discriminative	generative discriminative

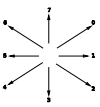
Applications

Original data | String representation | Tree representation | Learned parameters

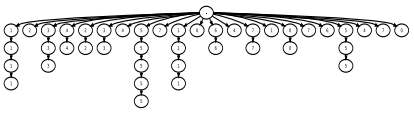
Handwritten Character Recognition

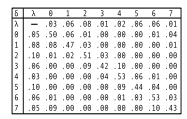


Use of Freeman's codes



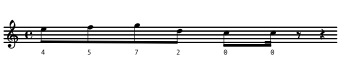
11112333442233455
55571111066477100
76555470



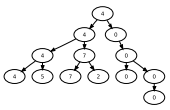


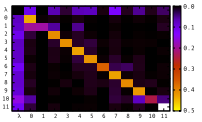
δ	λ	0	1	2	3	4	5	6	7
λ	—	.03	.06	.08	.01	.02	.06	.06	.01
0	.85	.50	.06	.01	.08	.00	.00	.01	.04
1	.08	.88	.47	.03	.08	.00	.00	.00	.01
2	.10	.01	.02	.51	.03	.00	.00	.00	.00
3	.06	.00	.00	.09	.42	.10	.00	.00	.00
4	.03	.00	.00	.00	.04	.53	.06	.01	.00
5	.10	.00	.00	.00	.00	.09	.44	.04	.00
6	.06	.01	.00	.00	.00	.01	.03	.53	.03
7	.05	.09	.00	.00	.00	.00	.00	.10	.43

Music Recognition



457200





PASCAL2 Pattern Analysis, Statistical Modelling and Computational Learning
CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
French ANR-05-MMSA-0016 Marmota project
Spanish research programme Consolider Ingenio-2010
Université Jean Monnet SAINT-ETIENNE
UNIVERSITÉ DE LYON
Universitat d'Alacant
Universidad de Alicante