A Contextual Normalised Edit Distance

Colin de la Higuera [1] and Luisa Micó [2]

[1]Laboratoire Hubert Curien
Université de Saint-Etienne
Colin.Delahiguera@univ-st-etienne.fr

[2]Dpto. Lenguajes y Sistemas Informáticos
Universidad de Alicante
mico@dlsi.ua.es

April 18, 2008

Summary

Summary

1. Introduction
2. Edit distance
3. The contextual edit distance
4. Experiments
5. Conclusions and future work

Introduction

- In Pattern Recognition, Computational Biology, Data Mining, Machine Learning ... there are some applications where data are represented by strings.

- The edit (Levenshtein) distance is a good candidate in many cases.

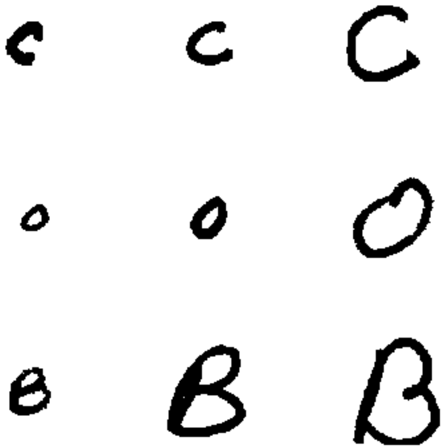
The problem

- Sometimes, the edit distance is not very suitable for some applications.
- Why? ... it lacks some type of normalisation

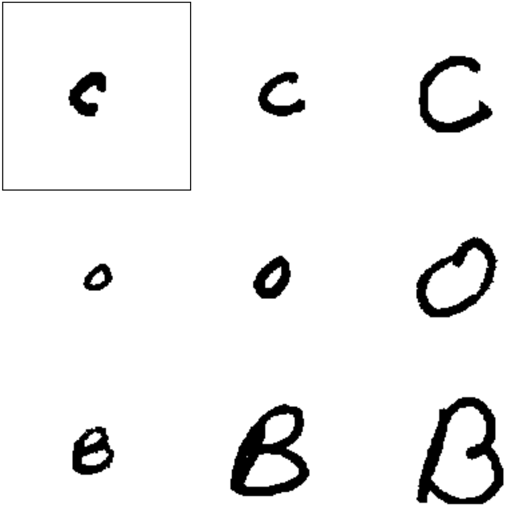
Examples:

- as ↔ on
- sun ↔ say
- performer ↔ peforme
- ornithological ↔ onitological
- supercalifragilisticexpialidocious ↔ supercalifragilisticoespialidocious

Example: handwritten character recognition



Example: handwritten character recognition



Example: handwritten character recognition

Edit distance

								
	52	62	61	115	94	62	156	157

Normalised edit distance

								
	0.51	0.74	0.54	0.71	0.63	0.45	0.87	0.88

Summary

Edit distance

Definitions

- An *alphabet* Σ is a finite nonempty set of symbols
- A *string* $x = x_1 \cdots x_n$ is any finite sequence of symbols
- Σ^* is the set of all the strings over Σ
- $|x|$ denotes the length of x

Edit distance

Given $x, y \in \Sigma^\star$, x **rewrites into** y **in** k **steps** ($x \xrightarrow{k} y$) using k operations of single symbol deletion, insertion or substitution.

The **edit distance** between x and y, $d_E(x, y)$, is the **smallest** k such that $x \xrightarrow{k} y$. [Levenshtein, 65]

Edit distance

Internal edit distance

The **internal edit distance**, $d_E^I(x, y)$, between x and y, is defined as the distance where only internal edit operations are allowed.

Then $d_E(x, y) = d_E^I(x, y)$

Example: $d_E(\overline{a}\overline{b}\overline{a}\overline{a}, \underline{baab}) \leq 3$ with an internal path

$$\overline{a}\overline{b}\overline{a}\overline{a} \rightarrow \underline{b}\overline{b}\overline{a}\overline{a} \rightarrow \underline{b}\overline{a}\overline{a} \xrightarrow{0} \underline{ba}\overline{a} \xrightarrow{0} \underline{baa} \rightarrow \underline{baab}$$

Length of the path $l_E(\pi) = 5$

Some normalised edit distances

- $d_{sum}(x, y) = \frac{d_E(x,y)}{|x|+|y|}$
- $d_{max}(x, y) = \frac{d_E(x,y)}{\max(|x|,|y|)}$
- $d_{min}(x, y) = \frac{d_E(x,y)}{\min(|x|,|y|)}$

more normalised edit distances

- $d_{MV}(x,y) = \min_{\pi}\left(\frac{d_E(\pi)}{l_E(\pi)}\right)$ [Marzal & Vidal, 1993]

- $d_{YB}(x,y) = \frac{2d_E(x,y)}{|x|+|y|+d_E(x,y)}$ [Yujian & Bo, 2007]

Objective of this work

Definition of a new normalised edit distance...

- that is a metric
- whose computational cost is small
- that has a good behaviour when using in fast NNS algorithms
- which works well in classification tasks

Summary

The contextual edit distance

Idea: consider that the weight of each edit operation is context dependent.

More precisely, when transforming u in v with an elementary operation, $d_C(u,v) = \frac{1}{max(|u|,|v|)}$

- substitution or deletion $\longrightarrow \frac{1}{|u|}$
- insertion $\longrightarrow \frac{1}{|u|+1}$

Example:

$$baabb \xrightarrow[\frac{1}{6}]{} bbaabb$$

Definition

Normalised contextual edit distance

The normalised contextual edit distance for a path $x = \omega_o \rightarrow \omega_1 \rightarrow ...\omega_k = y$ is $\sum_{i=1}^{k} d_c(\omega_{i-1}, \omega_i)$.

The normalised contextual edit distance between x and y is the **minimum** value $d_C(\pi)$ over all possible paths π from x to y.

Example: d(aabb, baa)

$$aabb \underset{\frac{1}{5}}{\longrightarrow} aabab \underset{\frac{1}{5}}{\longrightarrow} aabaa \underset{\frac{1}{5}}{\longrightarrow} abaa \underset{\frac{1}{4}}{\longrightarrow} baa$$

$$\text{d(aabb, baa)} = \frac{17}{20}$$

Properties

- the contextual edit distance is a metric
- $d_C(x, y) = d_C^I(x, y)$
- the best path for the contextual edit distance may not be optimal for the *usual* edit distance
- for a given length the best path maximises the number of insertions and first inserts, then substitutes and finally deletes

Key algorithmic idea

1. computing, for each value k, the maximum number $n_i(k)$ of insertions on a path of length k leading from x to y, and
2. finding the minimum value

$$\sum_{i=|x|+1}^{i=|x|+n_i(k)} \frac{1}{i} + n_s(k) \cdot \frac{1}{|x| + n_i(k)} + \sum_{i=|y|+1}^{i=|y|+n_d(k)} \frac{1}{i}$$

with

- $n_d(k) = |x| - |y| + n_i(k)$
- $n_s(k) = k - n_i(k) - n_d(k)$

Computational complexity

The complexity of the proposed algorithm is $O(|x| \cdot |y| \cdot (|x| + |y|))$

but

the minimum value is obtained very often for $k = d_E(x, y)$

This allows to consider a heuristic called $d_{C,h}$ which is in $O(|x| \cdot |y|)$

Summary

Experiments

- What about the intrinsic dimension?
- What about the behaviour with fast NNS algorithms?
- What about the error rate in classification tasks?

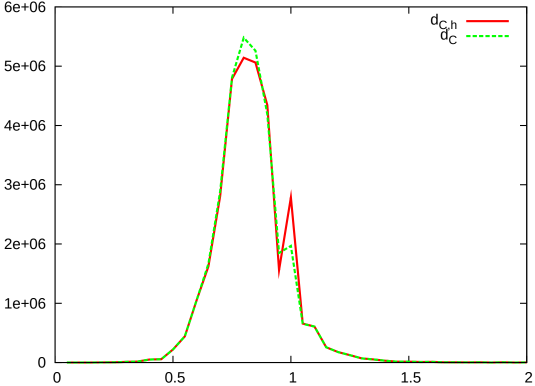
Datasets:

- a Spanish dictionary $\approx$ 80,000 words*
- a set of 20,000 DNA sequences of genes*
- a set of 10,000 contour strings of handwritten digits from the NIST Special Database 3

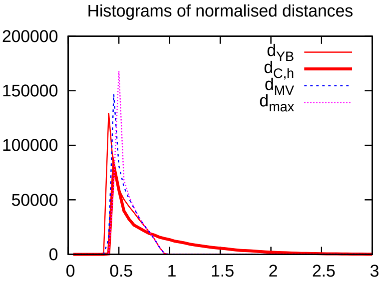
*from `http://sisap.org`

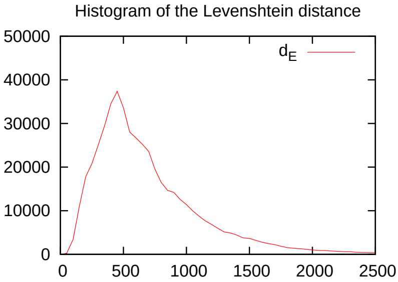
Using a heuristic

Comparison of d_C and $d_{C,h}$ for the Spanish dictionary (8000 samples)

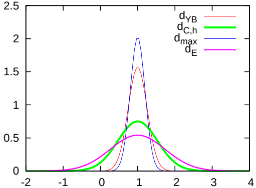


Dataset: genes

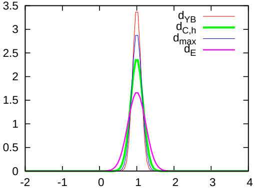




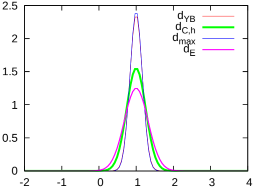
Normalisation



genes



Spanish d.



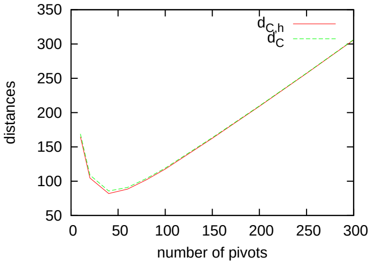
handwritten d.

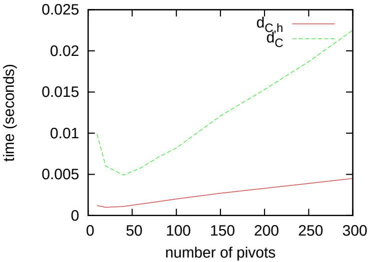
$\rho = \frac{\mu^2}{2\sigma^2}$ [Chávez et al, 2001]

Analysis of the intrinsic dimensionality

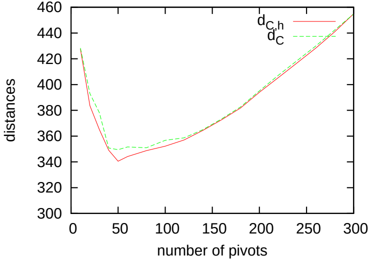
	Datasets		
Distances	Spanish D.	hand. digits	genes
d_{YB}	40.57	18.81	8.43
d_{MV}	33.98	19.36	11.25
d_{max}	30.25	19.48	14.13
$d_{C,h}$	**18.61**	**7.95**	**1.88**
d_E	8.75	4.91	0.99

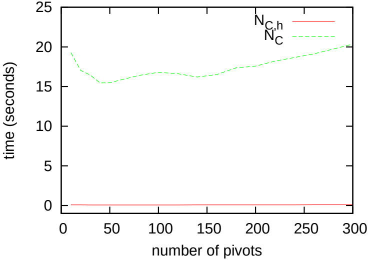
Comparison of d_C and $d_{C,h}$ for the Spanish dictionary (1000 samples)



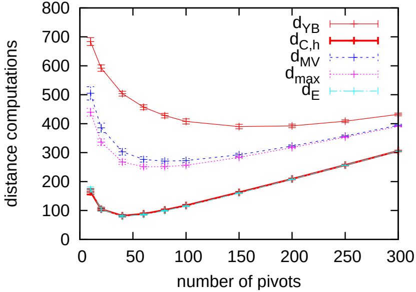


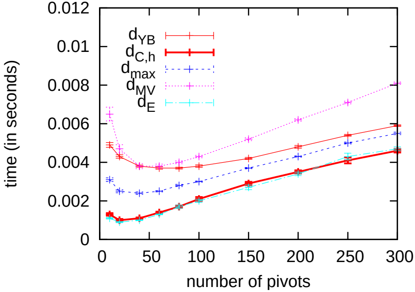
Comparison of d_C and $d_{C,h}$ for the handwritten character dataset (1000 samples)



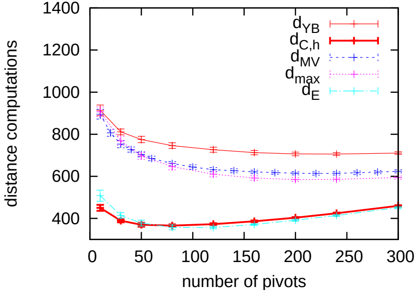


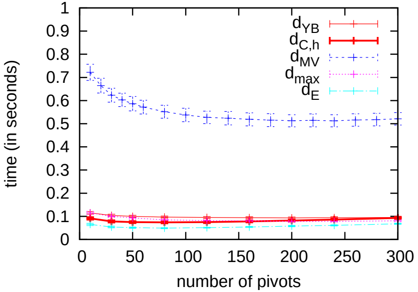
Dataset: Spanish dictionary





Dataset: Handwritten digits





Dataset: Handwritten digits

Distances	Error rate (%)
d_{YB}	5.19
d_{MV}	5.04
d_C	5.30
$d_{C,h}$	5.30
d_{max}	4.85
d_E	6.19

Summary

Conclusions

To summarise, we have proposed a new extension of the edit distance with the following properties:

- is a metric
- can be computed in cubic time, although an approximation is obtained in quadratic time;
- have a good behaviour when is used in fast NNS algorithms
- the error rate in a handwritten digit classification task is good

Future work

- further analysis is needed in order to reduce the complexity of the algorithm
- an adaptation of the technique to the generalised edit distance will be considered.