doi:10.4186/ej.2009.13.2.19

PSEUDOMETRICS FOR NEAREST NEIGHBOR CLASSIFICATION OF TIME SERIES DATA

Teesid Korsrilabutr* and Boonserm Kijsirikul**

Department of Computer Engineering,
Faculty of Engineering, Chulalongkorn University,
Bangkok, Thailand 10330
E-mail : *teesid@gmail.com, **boonserm.k@chula.ac.th

ABSTRACT

We propose that pseudometric, a subadditive distance measure, has sufficient properties to be a good structure to perform nearest neighbor pattern classification. There exist some theoretical results that asymptotically guarantee the classification accuracy of *k*-nearest neighbor when the sample size grows larger. These results hold true under the assumption that the distance measure is a metric. The results still hold for pseudometrics up to some technicality. Whether the results are valid for the non-subadditive distance measures is still left unanswered. Pseudometric is also practically appealing. Once we have a subadditive distance measure, the measure will have at least one significant advantage over the non-subadditive; one can directly plug such distance measure into systems which exploit the subadditivity to perform faster nearest neighbor search techniques.

This work focuses on pseudometrics for time series. We propose two frameworks for studying and designing subadditive distance measures and a few examples of distance measures resulting from the frameworks. One framework is more general than the other and can be used to tailor distances from the other framework to gain better classification performance. Experimental results of nearest neighbor classification of the designed pseudometrics in comparison with well-known existing distance measures including Dynamic Time Warping showed that the designed distance measures are practical for time series classification.

doi:10.4186/ej.2009.13.2.19

I . Introduction

Since its inception in the 1950s ([1], [2]), k -nearest neighbor (k -NN) still receives regular interest among researchers; both in the theoretical aspect and the practical aspect. Its discrimination procedure is simple but powerful and needs virtually no modification to handle multi-class problems, i.e. it just obeys the majority vote for the classes among the k nearest neighbors of the sample being considered. k -NN decision rules gained theoretical acceptance since its early age of development; [1] developed their notions of *consistencies* between sequences of decision functions and showed that a formulation of k -NN is consistent with a reference decision rule. Many notable points are worth mentioning in their work. They initiated the field of *nonparametric classification*, the distribution generating the examples need not be assumed to be Gaussian or any other parametric distributions. The reference decision rule mentioned in their work as the "likelihood ratio procedure" [3] is closely related to what is known today as the Bayes classifier. The Bayes classifier is the best classifier that will yield the lowest possible expected misclassification given that we know the distribution of the data; it will be discussed in detail later. They established that whenever the number n of available examples approaches infinity and k_n are dependent of n such that $k_n \to \infty$ and $k_n/n \to \infty$, the decision of k_n -NN will get arbitrarily closer to that of the *likelihood ratio procedure* with high probability. For example, one may choose k_n to be $\lceil \log n/n \rceil$. Later, [4] showed that simpler rules also possess good asymptotic properties; for a fixed k the error probability of k -NN will be at most twice that of the Bayes classifier in the limit as the number of examples grows to infinity. The link between k -NN error and Bayes error provides ways to estimate the theoretical limit one can achieve. Lower bounds of the Bayes error relative to errors of modified versions of k -NN rules are also studied.

More recently, various attempts to learn a good metric to use in k -NN classification have been proposed [5],[6],[7],[8]. Most of them learn the so called ``Mahalanobis distance'', which can be perceived as a Euclidean distance in a linear transformation of the original vector space of examples. Several objectives had been proposed and optimized in order to find the best linear transformation, and most of the proposed objectives are formulated to be able to be solved by convex optimization or the spectral method, where the optimum is guaranteed to be global. Some of them are optimized for local minimum by gradient descent algorithms or other non-convex optimization techniques. The common goal, however, is to optimize a quantity that are related to classification performance of k -NN; the learned metric is used in k -NN. In their experiments, k -NN with the learned metric even outperforms the current state of the art learners such as support vector machines for some datasets [6].

The naïve version of the k -NN algorithm is easy to implement by computing the distances from the test sample to all stored vectors, but it is computationally intensive, especially when the size of the training set is large. From the practical point of view, large scale k -NN classification scenarios face the problem of speed. Several distance measures are ideated to augment existing well known basic distances such as the Euclidean distance and the ℓ^p distances and in many cases the new distance measures outperform existing ones in terms of classification accuracy. However, more accurate distances come with their price; they usually need more time to compute. A well known example of such event is the DTW distance whose running time grows like the square of time series lengths, while the ℓ^p distance takes linear time of time series lengths. The distance measures in use nowadays may be classified exclusively into two different kinds, namely

1. subadditive distance measures: by definition a distance d is subadditive if $d(x,z) \le d(x,y) + d(y,z)$ for every x,y,z, and the inequality is called the triangle inequality or the triangle law,
2. non-subadditive distance measures, which is the complement of the first kind.

Subadditivity is useful in avoiding the need to compute every distance when the nearest neighbor is to be searched. A simple technique [9] to prune unnecessary computation of distance between some pair of items when doing nearest neighbor queries is to select an item from the pool of candidates which will be used as the *reference* item. The distance between the reference item and each of the candidates will be computed and stored in advance. Together with the distance between the reference item and the query item, those stored values can be used to lower bound the distance of the query item from each candidate item in constant time. If the lower bound distance from the query item to candidate

x is greater than the *closest so far* distance, then x can be safely abandoned without having to compute its distance from the query item. Various works that take the advantage of this fact exists ([9],[10],[11],[12]); most of the works were done by database researchers and can be used instantly if only the distance measure we use is subadditive.

Although the DTW distance cannot be lower bounded using the triangle inequality, one can compute the lower bound of the DTW distance between each pair of time series instead and such bounds can be similarly used to prune out futile computations of actual DTW distances. The best known strategy to lower bound the DTW distance is due to Keogh and Ratanamahatana [13]. Their lower bound can be computed in linear time.

A few questions arise naturally. Are the theoretical results regarding the asymptotic properties of k -NN applied for every distance in use today? Which of the widely used distances is of the first kind and which is not? What is a good distance for doing k -NN classification?

It may be unfair to the first question but we will answer the second first. Some widely used distances are not subadditive; examples are DTW and Shape Context Distance [14]. The distances that are of the first kind are the well known Euclidean and ℓ^p metrics, and instances of the less commonly known ones are Levenshtein distance or edit distance [15] and Edit Distance with Real Penalty (ERP) [16], for example.

The answer to the first question is, unfortunately, negative. All of the nice asymptotic results for k -NN require that the distance measure be either the Euclidean metric [1], a norm metric [17](chap. 5) or a metric with some assumptions [4]. Perhaps the least restrictive result, when considering only the conditions imposed on the distance used by k -NN, is in the work of [4], where the distance has to be a metric in a separable metric space, but since a non-subadditive distance fails to be a metric in the first place, k -NN with the second kind distances does not enjoy the existing results. Whether the results can be extended to cover non-metrics is still unknown. Although this does not necessarily imply that extensions of these nice results to non-metrics are impossible, it does indicate that more work has to be done in order to justify non-metrics k -NN theoretically. More precise statement regarding these asymptotic results will be formally given in Section 2.

As common sense and the formalized concept of ``no free lunch'' suggest [17](chap. 7), a good distance is inevitably dependent on the problem at hand. For the last question we will not try to give a clear cut answer. Instead, we give a partial answer by a list of desirable properties. For a given set of examples, if a distance measure has the following properties,

1. it is a pseudometric,
2. it gives good accuracy for the particular set of examples,

then we say that it is a good distance for doing k -NN, with respect to the examples. The exact definition of pseudometric will be given in Section 2. It is briefly a symmetric subadditive distance measure. The first property has twofold advantages. First, it ensures us, up to some assumptions, that our classifier has the potential to perform incrementally better when we have more observed examples in the future (a pseudometric can be regarded as a metric in a technically adjusted space). Second, subadditivity is useful to hasten nearest neighbor searches and we can plug a pseudometric into existing systems that take advantage of the triangle law if we want k -NN to be faster. So pseudometrics are both theoretically and practically salient. The existing asymptotic results, at least the work by [4], still hold for a pseudometric given that the underlying space is separable. The second property is vital in its own right.

Generalizing from Euclidean and ℓ^p spaces to metric and pseudometric spaces is somehow a sensible next step of development since metric spaces bear some relationship with ℓ^p spaces. Metric spaces are well studied. For example, it is well known that metric spaces are Hausdorff, implying that every convergent sequence has a unique limit, and any metric spaces can be embedded isometrically into a Banach space [18]. Several fixed point theories for metric spaces are in the mathematical literature [19]. Other than speed gains for the nearest neighbor algorithm, more interesting results may be discovered for pseudometrics k -NN or related algorithms as well.

This work is restricted to pseudometrics for univariate time series, although it will be seen that some results in our work hold for more abstract structures than just time series. We

attempt to study pseudometrics for time series first because time series are slightly different from vectors. We will be well equipped with tools and structures in linear spaces to work with.

In the remaining sections, we will make the problem setting more precise after the introduction of notations used throughout our expositions, followed by the main work corroborated with the experiments. Sufficient backgrounds and pointers to relevant references are in Section 2, one should be familiar with in order to follow the development. In Section 3, we introduce a concept called condensation to be used as a guideline for designing new distances. As a by product, we discover an alternative characterization of the DTW distance. The second distance construction guideline called ``shortcut distance'' will also be discussed in Section 3 and we will demonstrate how it can be used to fine tune distances to yield better empirical classification performance. Numerical results are in Section 4. Conclusion and future work are given in Section 5.

II. Background

2.1 Conventions

Random variables are uppercase characters such as X, Y, and Z. We usually denote a time series or a finite dimensional vector by a boldface letter such as $\mathbf{s}$, and its length by $\#\mathbf{s}$ or just l when the mentioned time series is obvious. The i-th value of the time series $\mathbf{s}$ is written s_i or $\mathbf{s}\langle i\rangle$, and by $\mathbf{s}$ and $[s_1,\ldots,s_l]$ we mean the same thing.

The time series $[s_2,\ldots,s_l]$ is called the *tail* of $\mathbf{s}$, denoted by $\mathbf{s}_{\sim}$. $\mathbf{0}$ and $[]$ are $[0]$ and the time series of length zero, respectively. There is one and only one place, in Section 2.4, where the square brackets enclosing a letter [x] will be used to denote the equivalence class of x and we do not mention time series there.

Functions that change a vector or time series or one object to another, called *morphs*, are denoted by Greek letters using prefix notation. For example, $\mu(x)$ or μx is understood as the morphed object from x by the morph μ. Compositions of functions such as $\mu(\nu(x))$ may be written as $\mu\nu x$ or $\mu\circ\nu x$. $\mathbb{I}$ denotes the identity map.

Calligraphic scripts such as $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{M}$ are used to denote sets of functions. For a set $\mathcal{F}$ of functions on a space Ω and $x\in\Omega$, we let $\mathcal{F}(x)$ be the set $\{f(x)\,|\,f\in\mathcal{F}\}$.

We denote infinite sequences by the list of its elements enclosed in a parentheses e.g. $(1,2,3,4,\ldots)$ Depending on the context, sometimes we regard a finite sequence as an infinite sequence entailed with zeros, or as an infinite sequence entailed with a constant sequence of its last element. For example, we may think of [0.5,1] as $(0.5,1,0,0,0,\ldots)$ or $(0.5,1,1,1,\ldots)$, subject to the context.

Given a finite sequence $\mathbf{s}$ and another sequence $\mathbf{t}$, the concatenation of $\mathbf{s}$ and $\mathbf{t}$ is written as $\mathbf{st}$, and $\mathbf{s}^1$ is the same as $\mathbf{s}$ and $\mathbf{s}^n$ is defined recursively as $\mathbf{ss}^{n-1}$.

2.2 Classification Problem Settings

We follow the same setting as in the work of [17]. The c-class classification problem in a probabilistic setting is formalized as follows. Let (X,Y) be a pair of random variables taking values in the Cartesian product $\Omega\times\Lambda$ of a metric space Ω of all possible examples and the class labels $\Lambda=\{1,\ldots,c\}$. A function $f:\Omega\to\Lambda$ deciding the class label based solely on the observation of examples from Ω is called a *classifier*. A *rule*, upon a given finite set of i.i.d. pairs of values observed from the random pair (X,Y), constructs a classifier. For example, in the case of 1-NN rule, given $\{(x_1,y_1),\ldots,(x_n,y_n)\}$, it constructs the decision function

doi:10.4186/ej.2009.13.2.19

$$g_n(x) = y_k,$$

where x_k is closest to x.

An error occurs if $f(X) \neq Y$, and the probability of error for a classifier f is $L(f) = \mathbf{P}\{f(X) \neq Y\}$.

For a fixed rule, the classifier f_n constructed according to n observations from the random pair (X,Y) depends randomly on the data sequence, so as the conditional probability of error

$$L_n = L(f_n) = \mathbf{P}\{f_n(X) \neq Y \mid X_1, Y_1, \ldots, X_n, Y_n\}.$$

2.3 Bayes Classifier

The Bayes classifier is the following decision function

$$g^*(x) = \underset{i \in \Lambda}{\operatorname{argmax}} \mathbf{P}\{Y = 1 \mid X = x\}.$$

It can be shown [17](chap. 2) that for any classifier g,

$$L^* = \mathbf{P}\{g^*(X) \neq Y\} \leq \mathbf{P}\{g(X) \neq Y\}.$$

2.4 Distance, Metric and Norm

The asymptotic results of k-NN contains different assumptions on the distance measure and the probability distribution of the data. In order to get a good grasp of the different assumptions on the distance measures, we begin with the definition of metric space and its relatives.

Definition 1. *A metric space* (Ω, d) *is a set* Ω *together with a non-negative extended-real-valued function* $d : \Omega \times \Omega \to [0, \infty]$ *(called a metric) such that, for every* $x, y, z \in \Omega$,

(1) $d(x,x) = 0$,
(2) $d(x,y) = 0$ *implies* $x = y$,
(3) $d(x,y) = d(y,x)$,
(4) $d(x,z) \leq d(x,y) + d(y,z)$.

If the condition (2) is omitted then we have a pseudometric space *and d will be called a* pseudometric*. If the conditions (2) and (4) are dropped, we have a* distance space *and call d a* distance*.*

Note that we allow d to take the value ∞ so that the definition of a metric and its relatives are technically applicable in a wider situation. For example, if some pair of elements in Ω are not comparable, we let their distance be ∞.

Pseudometric space is a salient structure to perform nearest neighbor queries. For a pseudometric space, numerous techniques [9],[10],[11],[12],[20] could be readily applied to speed up nearest neighbor queries, and sometimes k-means algorithms. Those works are based on the following bounding scheme or their variants, each of them is derivable from the triangle law,

$$d(x,z) \geq |\, d(x,y) - d(y,z) \,|,$$

$$d(x,y) \geq \frac{1}{2} d(x,z) - |\, d(y,z) - \frac{1}{2} d(x,z) \,|.$$

Examples of pseudometric spaces are, the real numbers with absolute difference, vector spaces with the Euclidean distance, a set of strings with edit distance, etc.
Having a pseudometric space (Ω,ρ) with a pseudometric ρ, we can always have a metric space by gluing together elements in Ω that are zero distance apart together. Precisely, we construct a *quotient space* of it using the equivalence relation

$$x \sim y \quad \text{iff} \quad \rho(x,y) = 0.$$

Then we can think of an equivalence class as one point in the new space and define a new metric in the space $(\Omega/\sim)$ of equivalence classes,

$$\tilde{\rho}([x],[y]) = \rho(x,y).$$

The new space $\Omega/\sim$ is a metric space.

It is not uncommon that one encounters norms when working with vector spaces. Since we may think of a set of time series as a vector space of number sequences, norms are involved naturally. Some of the asymptotic results for k -NN hold for norms metrics. They will also be mentioned later in Section 3.

Definition 2. *Let V be a vector space. A function $\|\cdot\|: V \to [0,\infty)$ is a norm on V iff for every $x,y \in V$*

(1) $\| x \| = 0$ iff $x = 0$,
(2) $\| \alpha x \| = | \alpha | \| x \|$ for any $\alpha \in \mathbb{R}$,
(3) $\| x + y \| \leq \| x \| + \| y \|$.

Having a normed space, the function $(x,y) \mapsto \| x - y \|$ is always a metric. Well known norms for the space of number sequences are the ℓ^p norms defined by $\| \mathbf{x} \|_p = \{\sum_{i=1}^{l} | x_i |^p\}^{\frac{1}{p}}$, for $p \in [1,\infty)$ and $\| \mathbf{x} \|_\infty = \max_i | x_i |$. The latter is called the *supremum norm*. The metrics induced by ℓ^p norms are called ℓ^p metrics.

2.5 Asymptotic Behavior of Metric Based k-NN

In terms of generality, there are two major asymptotic results; the first holds for separable metric spaces but with a usual assumptions on the distribution.

Theorem 1 [4]. *Let Ω be a separable metric space and (X,Y) admits class conditional densities. Let $f_1,\ldots,f_c$ be probability densities such that $f_i(x) = \mathbf{P}\{X = x \mid Y = y\}$ and f_i is continuous almost everywhere for each $i \in \{1,\ldots,c\}$. Then the k -NN probability of error L has the bounds*

$$L^* \leq \lim_{n\to\infty} \mathbf{E} L_n \leq L^* \left(2 - \frac{cL^*}{c-1} \right) \leq 2L^*.$$

These bounds are as tight as possible.

The second major result holds for every possible distribution but the distance is assumed to be a norm metric [17](prob. 5.1, chap. 5).

Theorem 2 [17](chap. 5)**.** *Let the random pair (X,Y) take values in $\mathbb{R}^d \times \{1,2\}$. Then the norm-metric based k -NN probability of error L has the bounds*

$$L^* \leq \lim_{n\to\infty} \mathbf{E} L_n \leq L^* \left(2 - 2L^* \right) \leq 2L^*.$$

doi:10.4186/ej.2009.13.2.19

Since our pseudometrics in Section 3 are not norm metrics we have to hinge on the former theorem and assume the regularity of the distributions. Note that a quick argument that the set of all time series with all rational values serves as a countable dense subset will be sufficient to establish that all of the pseudometric spaces of time series in Section 3 are separable. Such argument ensures that the space of time series equipped with DTW as the distance is a separable distance space, although not a metric space.

2.6 Admissibility of 1-NN

[4] showed in their paper that 1-NN is admissible in the sense that there is a distribution of data such that k -NN will be strictly worse than 1-NN in terms of probability of misclassification for every $k>1$ .They also give an example of such distribution in the paper and noted that if the between-class distances are always greater than the within-class distances then1-NN is strictly better than any other k -NN.

2.7 DTW Distances

In computing the DTW distance, one searches for a *warping path* with the lowest possible associated cost. Given two finite sequences of real numbers **s** and **t** , a warping path between **s** and **t** is a sequence of pairs

$$(i_1, j_1),(i_2, j_2),\ldots,(i_N, j_N) \qquad (1)$$

Where $N \le \#\mathbf{s} + \#\mathbf{t} - 1$, $(i_1, j_1) = (1,1)$, $(i_N, j_N) = (\#\mathbf{s}, \#\mathbf{t})$, and (i_k, j_k) must be one of $(i_{k-1}, j_{k-1}),(i_{k-1}, j_k)$ or (i_k, j_{k-1}) for all $2 \le k \le N$.

One may perceive a warping path as a continuous monotonic sequence of coordinates whose start and end are fixed in a two dimensional grid. The cost associated with a warping path in Equation (1) is $\sum_{k=1}^{N} d(s_{i_k} - t_{j_k})$, where d is any distance measure – common choices are absolute difference and squared difference.

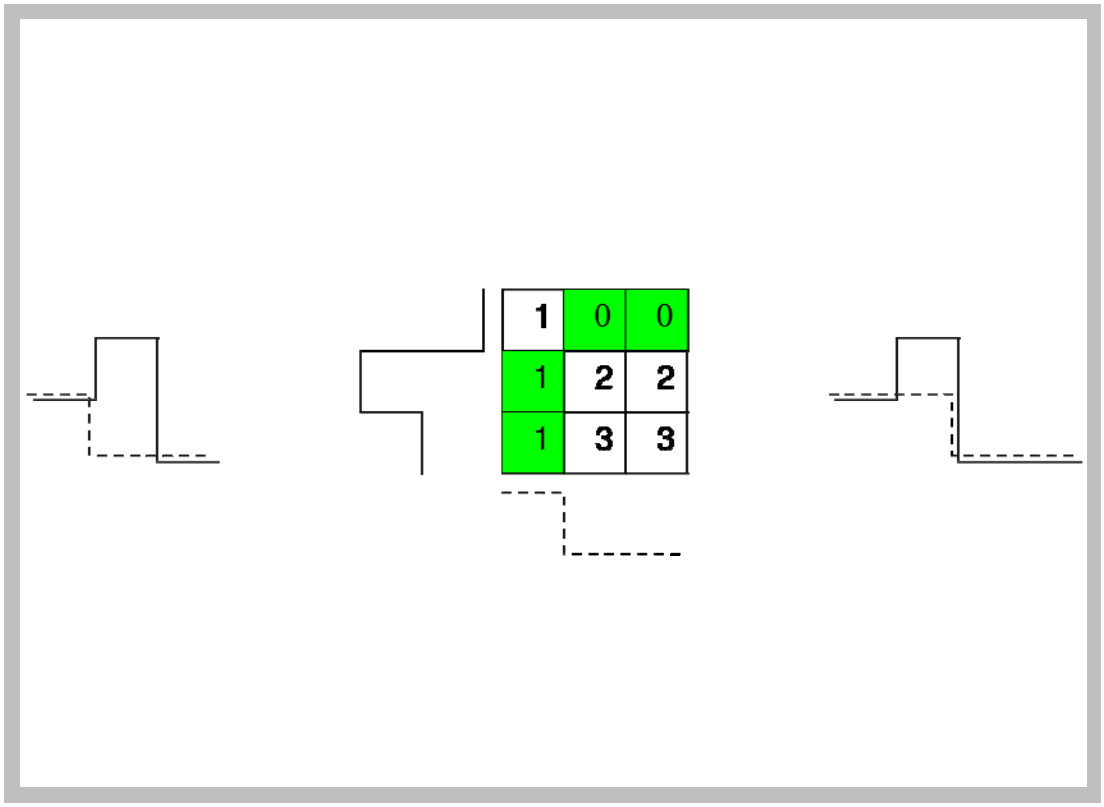

Figure 1
A visualization of a warping path, which is also an optimal warping path whose associated cost is 1. The path is $(1,1),(2,1),(3,2),(3,3)$ from bottom left to top right. On the left, the dotted line is the sequence $[1,2,0]$ and other in solid line is the sequence $[1,0,0]$. On the right, the dotted line is the sequence $[1,2,0,0]$, which is a *stretch* (see Definition 6) of $[1,2,0]$; i.e. $[1,2,0,0] \in \mathcal{S}([1,2,0])$. The solid line on the right is $[1,1,0,0]$, which is a stretch of $[1,1,0]$. Note that the distance between $[1,2,0,0]$ and $[1,1,0,0]$ is equal to the cost of the optimal warping path.

Figure 1 shows an example of the optimal warping path of two time series $[1,0,0]$ and $[1,2,0]$; the path is $(1,1),(2,1),(3,2),(3,3)$ from the bottom left of the grid to top right. Suppose for concreteness that d is absolute difference. The DTW distance can be expressed in terms of its partial solutions as,

$$\mathrm{DTW}(\mathbf{s},\mathbf{t}) = |\, s_1 - t_1 \,|^p + \min\begin{cases}\mathrm{DTW}(\mathbf{s}_{\sim},\mathbf{t}_{\sim}),\\ \mathrm{DTW}(\mathbf{s},\mathbf{t}_{\sim}),\\ \mathrm{DTW}(\mathbf{s}_{\sim},\mathbf{t}).\end{cases}$$

Where $\mathrm{DTW}([s_1],\mathbf{t}) = \mathrm{DTW}(\mathbf{t},[s_1]) = \sum_{i=1}^{\#\mathbf{t}} |\, s_1 - t_i \,|^p$ for the base cases and p is usually 1 or 2 as mentioned above. $\mathrm{DTW}(\mathbf{s},\mathbf{t})$ can be computed in $O(\#\mathbf{s}\#\mathbf{t})$ time.

```
DTW-Distance(A[1..n],B[1..m])
    W[0][0] ← 0
    W[0][1..m],W[1..n][0] ← ∞
    for i ← 1 to n
        do for i ← 1 to m
            do d ← min{W[i][j − 1],W[i − 1][j],W[i − 1][j − 1]}
                W[i][j] ←| A[i] − B[j] |^p + d
    return W[n][m]
```

Figure 2
Pseudocode of the DTW algorithm.

By the expression above, the DTW distance can be computed by dynamic programming paradigm. Pseudocode of the DTW algorithm is shown in Figure 2. More detailed treatment of the subject can be found in other sources [21],[22],[23],[24].

2.7.1 Non-Subadditivity of DTW

To see that the DTW distance is not a pseudometric, consider the following trivial example. Let $\mathbf{s} = [1,1]$ and $\mathbf{t} = [1,1,1]$, then $\mathrm{DTW}(\mathbf{0},\mathbf{s}) + \mathrm{DTW}(\mathbf{s},\mathbf{t}) = 2 + 0 = 2$, while $\mathrm{DTW}(\mathbf{0},\mathbf{t}) = 3$. Another interesting example is when $\mathbf{u} = (1,0,0), \mathbf{v} = (1,2,0)$. We have $\mathrm{DTW}(\mathbf{0},\mathbf{v}) = 3 > 2 = \mathrm{DTW}(\mathbf{0},\mathbf{u}) + \mathrm{DTW}(\mathbf{u},\mathbf{v})$. These demonstrate that the DTW distance is not subadditive, and hence not a pseudometric.

2.8 Levenshtein Distance

The Levenshtein distance is a metric used to measure difference between two strings. The following relation may be taken as its definition

$$\mathrm{Lev}(\mathbf{s},\mathbf{t}) = \rho(s_1,t_1) + \min\begin{cases}\mathrm{Lev}(\mathbf{s}_{\sim},\mathbf{t}_{\sim}),\\ \mathrm{Lev}(\mathbf{s},\mathbf{t}_{\sim}),\\ \mathrm{Lev}(\mathbf{s}_{\sim},\mathbf{t}).\end{cases}$$

Where the function $\rho(x,y)$ is the *discrete metric* taking value 0 if x equals y and 1 otherwise. $\mathrm{Lev}([\,],[\,]) = 0$ and $\mathrm{Lev}([\,],[s_1]) = 1$ for the base cases.

The Levenshtein distance between two strings is the minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion, or substitution of a single character.

The distance is subadditive, indeed it is a metric, and one way to see this is by the fact that the distance is the minimum number of operations needed to transform one string

doi:10.4186/ej.2009.13.2.19

to the other. For strings $\mathbf{s}$, $\mathbf{t}$ and $\mathbf{u}$, the sum $\mathrm{Lev}(\mathbf{s},\mathbf{u})+\mathrm{Lev}(\mathbf{u},\mathbf{t})$ is the number of an operation sequence that transforms $\mathbf{s}$ to $\mathbf{t}$ (by changing $\mathbf{s}$ to $\mathbf{u}$ and then to $\mathbf{t}$), but that number is never greater than $\mathrm{Lev}(\mathbf{s},\mathbf{t})$ which is the minimum the length of such operations.

2.9 Edit Distance with Real Penalty

Edit Distance with Real Penalty (ERP) [16] is adapted from the Levenshtein distance. It is subadditive via a result for edit distance by [25].

For a real valued γ called ``gap'' ERP is defined by

$$\mathrm{ERP}(\mathbf{s},\mathbf{t}) = \min\begin{cases} |s_1 - t_1| + \mathrm{ERP}(\mathbf{s}_{\sim},\mathbf{t}_{\sim}), \\ |\gamma - t_1| + \mathrm{ERP}(\mathbf{s},\mathbf{t}_{\sim}), \\ |s_1 - \gamma| + \mathrm{ERP}(\mathbf{s}_{\sim},\mathbf{t}). \end{cases} \quad (2)$$

Where $\mathrm{ERP}([],\mathbf{s}) = \mathrm{ERP}(\mathbf{s},[]) = \sum_{i=1}^{l} |\gamma - s_i|$ for the base cases.

The value γ of the gap can be thought of as the default value in the sense that the constant sequence of γ, $(\gamma,\gamma,\ldots)$ is the null signal. It usually makes sense that the gap value is set to zero in practice because we usually perceive the null signal as a sequence of zeros. A side benefit is that we do not need to compute the difference of the gap value and the element of another sequence if the gap is zero.

III. Pseudometrics for Time Series

In this section we propose two guidelines and examples of applications of the guidelines.

3.1 Condensations of Distances

First of all, our notion of distance condensation should not be confused with the concept condensing by [17](chap. 19), where the data points are eliminated such that the classification is kept unchanged. Having a set of structured data, we usually have a simple distance measure that is easy to compute but gives undesirable classification accuracy when used with the nearest neighbor algorithm. Sometimes we want to allow variations of the two objects in a controllable manner so that they become more similar before we decide how different the two objects are. For example, two signals whose shapes are almost the same but one arrives a second later than the other should be considered almost the same without taking the time shift into account.

With a set of *morphs* allowed to be made to objects before being compared, we can always define another distance function.

Definition 3. *Let Ω be a distance space with the distance d, and $\mathcal{M}$ be a set of functions from Ω to Ω. The distance*

$$\underset{d,\mathcal{M}}{\Delta}(x,y) := \inf_{\mu,\nu\in\mathcal{M}} d(\mu x,\nu y),$$

is called the condensation of d with respect to $\mathcal{M}$.

Note that the value of $\Delta_{d,\mathcal{M}}$ is never greater than d. The idea of condensation is depicted in Figure 3; one may think that each point x in the space Ω is mapped to $\{\mu x \mid \mu \in \mathcal{M}\}$, the set of its all possible morphs by functions in $\mathcal{M}$, and the new distance is the distance between such sets (the distance between sets in this sense is the distance between their closest elements).

doi:10.4186/ej.2009.13.2.19

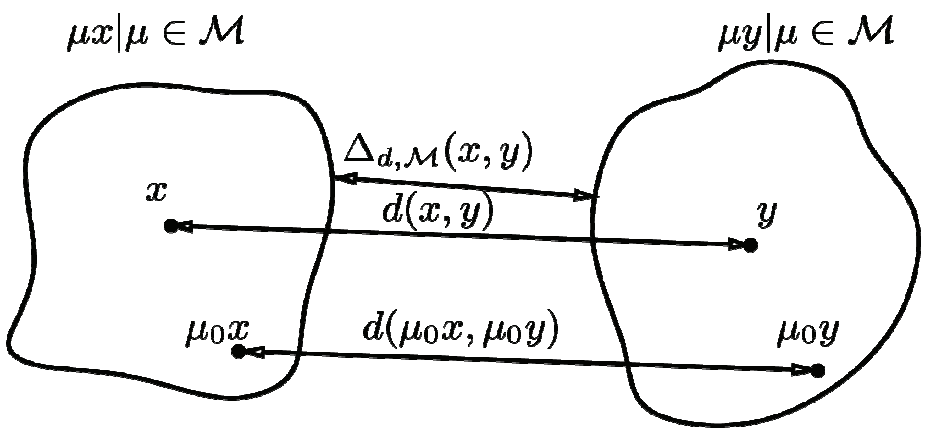

Figure 3
The condensation Δ_d measured between points x and y when $\mathcal{M}$ preserves d, μ_0 is a function in $\mathcal{M}$.

For a set of objects, two main components constitute a good distance measure; a base pseudometric and a set of *morph* operations of the objects with desirable properties. We give the definition of such morph operations below.

Definition 4. *Let $\mathcal{M}$ be a set of functions from Ω to Ω. $\mathcal{M}$ is said to be complete when,*

(1) the identity map is in $\mathcal{M}$,
(2) for each $\mu,\nu \in \mathcal{M}$, the composition $\nu\mu$ is in $\mathcal{M}$,
(3) for each $\mu_1,\nu_1 \in \mathcal{M}$, there are $\mu_2,\nu_2 \in \mathcal{M}$ such that $\mu_2\mu_1 = \nu_2\nu_1$.

We write an application of a function μ in $\mathcal{M}$ to an element x using the prefix notation μx. Compositions are read from right to left i.e. $\mu_2\mu_1 x$ is the result of an application of μ_2 to $\mu_1 x$.

Condition (3) in the definition above is weaker than the requirement that the composition of functions in $\mathcal{M}$ is commutative, i.e. $\nu\mu = \mu\nu$ for every μ,ν in $\mathcal{M}$. It is also weaker than requiring that every function in $\mathcal{M}$ has an inverse. So if a set of functions over Ω is a group, it is always complete in this sense.

Intuitively, with a complete set of morphs, two objects morphed from the same object remains similar, in a sense that they are the same up to some further morphing. Time shift is an example of a complete set of operations.

The following definitions are based on how a whole set of objects change their distance among each other when an operation is applied to the whole set.

Definition 5. *Let Ω be a pseudometric space equipped with a pseudometric d, and $\mathcal{M}$ be a set of functions from Ω to itself. We say that $\mathcal{M}$ preserves d if and only if,*

$$\forall x,y \in \Omega \quad \forall \mu \in \mathcal{M} \qquad d(\mu x, \mu y) = d(x,y),$$

$\mathcal{M}$ contracts d if and only if,

$$\forall x,y \in \Omega \quad \forall \mu \in \mathcal{M} \qquad d(\mu x, \mu y) \le d(x,y),$$

$\mathcal{M}$ expands d if and only if,

$$\forall x,y \in \Omega \quad \forall \mu \in \mathcal{M} \qquad d(\mu x, \mu y) \ge d(x,y).$$

We may say that $\mathcal{M}$ is contractive or expansive when the associated distance is implicitly known.

doi:10.4186/ej.2009.13.2.19

In Figure 3 we illustrate that μ_0 preserves the distance d. One may perceive d as the spatial distance on the paper and μ_0 as the translation by a certain amount to the southeast direction, and translating two points at the same time keeps their distance.
It turns out that if we want a subadditive condensation distance wrt.complete morphs, we should focus our interest on contractive ones.

Theorem 3. *Let* (Ω,d) *be a pseudometric space and* $\mathcal{M}$ *be a complete set of morph operations on* Ω. *Then the condensation of* d *wrt.* $\mathcal{M}$ *is a pseudometric if* $\mathcal{M}$ *contracts* d.

Proof. For brevity we write Δ to denote the condensation of d wrt. $\mathcal{M}$ throughout the proof.

Assume that (Ω,d) is a pseudometric space and $\mathcal{M}$ is a complete set of morph operations on contracting d. Obviously, $\Delta(x,x)=0$ for every x in Ω. The symmetry of Δ follows from the symmetry of d. Let $x,y,z\in\Omega$, $\varepsilon>0$. By definition of Δ there are some $\mu_1,\mu_2,\nu_1,\nu_2\in\mathcal{M}$ such that the following hold,

$$d(\mu_1 x,\mu_2 y)<\Delta(x,y)+\frac{\varepsilon}{2}, \tag{3}$$

$$d(\nu_1 y,\nu_2 z)<\Delta(y,z)+\frac{\varepsilon}{2}. \tag{4}$$

Since $\mathcal{M}$ is complete , there are $\mu_0,\nu_0\in\mathcal{M}$ making $\mu_0\mu_2 y=\nu_0\nu_1 y$. By definition,

$$\begin{aligned}\Delta(x,z) &\le d(\mu_0\mu_1 x,\nu_0\nu_2 z)\\ &\le d(\mu_0\mu_1 x,\mu_0\mu_2 y)+d(\nu_0\nu_1 y,\nu_0\nu_2 z)\\ &\le d(\mu_1 x,\mu_2 y)+d(\nu_1 y,\nu_2 z)\\ &< \Delta(x,y)+\Delta(y,z)+\varepsilon.\end{aligned}$$

The second inequality follows from $\mu_0\mu_2 y=\nu_0\nu_1 y$, the equality is by the assumption that $\mathcal{M}$ contracts d, and the last inequality follows from (3) and (4).

This is true for arbitrary $\varepsilon>0$, so $\Delta(x,z)\le\Delta(x,y)+\Delta(y,z)$.

3.1.1 Examples

A simple and trivial example of this kind of pseudometrics is the condensation of the Euclidean distance in a vector space wrt. arbitrary rotations about the origin. The resulting metric is just the difference between the lengths of its two arguments.

As another example, we construct a condensation of the ℓ^∞ metric wrt. the *stretch* operations defined below

Definition 6. *Let* V *be the set of all finite sequences. For each* $k\in\mathbb{N},\mathbf{x}\in V$ *define* $\sigma_k:V\to V$ *by,*

$$\sigma_k\mathbf{x}=\begin{cases}[x_1,\ldots,x_{k-1},x_k,x_k,x_{k+1},\ldots,x_l] & \text{if } k<l,\\ [x_l,\ldots,x_{k-1},x_l,x_l] & \text{if } k=l,\\ \mathbf{x} & \textit{otherwise}.\end{cases}$$

Let $\mathcal{S}$ *be the set containing every finite compositions of morph operations in* $\{\sigma_k\}_{k\in\mathbb{N}}\cup\{\mathbb{I}\}$.

One can verify that $\mathcal{S}$ preserves ℓ^∞ metric. Figure 4 shows how stretches of a time series may look like.

doi:10.4186/ej.2009.13.2.19

As a consequence, the following function is a pseudometric on V ,

$$\delta_1(\mathbf{x},\mathbf{y}) = \inf_{\mu,\nu\in\mathcal{S}} \| \mu\mathbf{x} - \nu\mathbf{y} \|_\infty . \qquad (5)$$

Where $\| \mathbf{x} - \mathbf{y} \|_\infty$ is implicitly defined to be ∞ when $\#\mathbf{x} \neq \#\mathbf{y}$.

Similar to DTW, the distance above can be written in terms of its partial answers.

$$\delta_1(\mathbf{s},\mathbf{t}) = \max\left(|s_1 - t_1|, \min\begin{Bmatrix} \delta_1(\mathbf{s}_\sim,\mathbf{t}_\sim), \\ \delta_1(\mathbf{s}_\sim,\mathbf{t}), \\ \delta_1(\mathbf{s},\mathbf{t}_\sim) \end{Bmatrix} \right), \qquad (6)$$

where $\delta_1([s_1],\mathbf{t}) = \delta_1(\mathbf{t},[s_1]) = \max\{|s_1 - t_1|,\ldots,|s_1 - t_i|\}$ for the initial cases. $\delta_1(\mathbf{s},\mathbf{t})$ can be computed in $O(\#\mathbf{s}\#\mathbf{t})$.

By Equation (6), one can be convinced that the distance function above is computed by the algorithm in Figure 5.

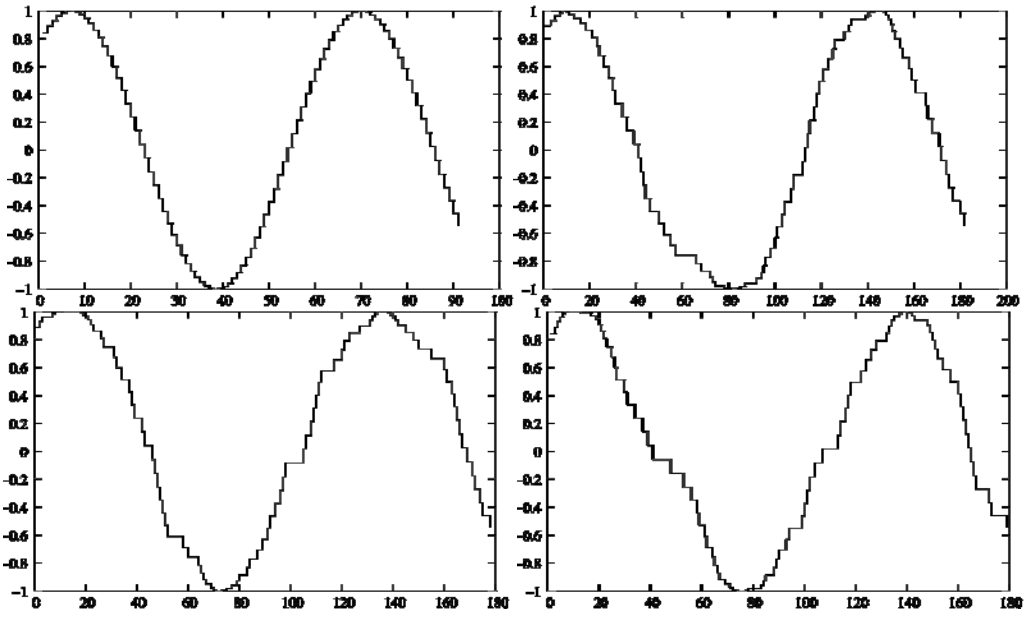

Figure 4
A visualization of a time series and its possible stretches. The original time series of length 91 is the top-left. The rest are some of its possible stretches to twice the original length.

```
Delta1(S[1..n],T[1..m])
    W[0][0] ← | S[1] − T[1] |
    W[0][1..m],W[1..n][0] ← ∞
    for i ← 1 to n
        do for i ← 1 to m
            do μ ← min{W[i − 1][j],W[i − 1][j − 1],W[i][j − 1]}
                W[i][j] ← max{| S[i] − T[j] |,μ}
    return W[n][m]
```

Figure 5
The algorithm computing δ_1

doi:10.4186/ej.2009.13.2.19

The following lemma will aid our further discussion. Via the lemma we will show, by a proof sketch motivated by examples, that the quantity in Equation (5) is equal to the quantity defined recursively in Equation (6). The proof of the lemma will be deferred until we end the proof sketch.

Lemma 1. *Let* $\mathbf{s}$ *and* $\mathbf{t}$ *be two finite sequences. If* $\mathbf{s}'$ *and* $\mathbf{t}'$ *are stretches of* $\mathbf{s}$ *and* $\mathbf{t}$ *respectively, i.e.* $\mathbf{s}' \in \mathcal{S}(\mathbf{s})$ *and* $\mathbf{t}' \in \mathcal{S}(\mathbf{t})$ *, such that* $\#\mathbf{s}' = \#\mathbf{t}' > \#\mathbf{s} + \#\mathbf{t} - 1$ *. Then there are* $\mathbf{s}''$ *and* $\mathbf{t}''$ *such that* $\mathbf{s}' \in \mathcal{S}(\mathbf{s})$ *,* $\mathbf{t}'' \in \mathcal{S}(\mathbf{t})$ *,* $\#\mathbf{s}'' = \#\mathbf{t}'' < \#\mathbf{s}' = \#\mathbf{t}'$ *and* $\| \mathbf{s}'' - \mathbf{t}'' \|_p \le \| \mathbf{s}' - \mathbf{t}' \|_p$ *. Where* $p \in [1, \infty]$ *.*

The two quantities in Equations (5) and (6) will be called the LHS and the RHS respectively.

1) Similar to DTW, after solving for the RHS, we have an optimal *warping path*,

$$(i_1, j_1), (i_2, j_2), \ldots, (i_N, j_N). \tag{7}$$

The restrictions of the optimal path is the same. The only difference is the associated cost. In this case the cost associated with the warping path is

$$\max_{k \in \{1,\ldots,N\}} | s_{i_k} - t_{j_k} | .$$

2) From the warping path we can construct two *stretched* time series whose lengths are equal and not greater than $\#\mathbf{s} + \#\mathbf{t} - 1$. Furthermore, their distance as computed by ℓ^∞ metric is equal to its associated cost. The two time series constructed are $[s_{i_1}, \ldots, s_{i_N}]$ and $[t_{j_1}, \ldots, t_{j_N}]$. For example, from the warping path in Figure 1 we construct $[1,1,0,0]$ and $[1,2,0,0]$, which are stretches of $[1,0,0]$ and $[1,2,0]$ respectively.

3) Conversely, for each pair of *stretched* $\mathbf{s}$ and $\mathbf{t}$ whose lengths are equal and do not exceed $\#\mathbf{s} + \#\mathbf{t} - 1$, we can construct a warping path with the associated cost equal to the ℓ^∞ distance between those two stretched time series. For example, suppose $\mathbf{s}$ and $\mathbf{t}$ are $[1,2,0]$ and $[1,0,0]$, the stretched time series $\mathbf{s}'$ and $\mathbf{t}'$ are $[1,2,0,0]$ and $[1,1,0,0]$. Noticing that $\mathbf{s}'$ and $\mathbf{t}'$ are $[s_1, s_2, s_3, s_3]$ and $[t_1, t_1, t_2, t_3]$, we can construct the warping path $(1,1),(2,1),(3,2),(3,3)$ having the desired associated cost.

4) Therefore the set of all possible costs associated with a warping path of $\mathbf{s}$ and $\mathbf{t}$ is the same as the set of all possible ℓ^∞ distance between $\mathbf{s}' \in \mathcal{S}(\mathbf{s})$ and $\mathbf{t}' \in \mathcal{S}(\mathbf{t})$ of equal lengths not exceeding $\#\mathbf{s} + \#\mathbf{t} - 1$. We conclude that RHS can be viewed as the minimum cost among all distances of such pair of stretched time series.

5) So by Lemma 1, we have LHS $\geq$ RHS. On the other hand, since LHS is an infimum taken over bigger set than that of RHS, LHS $\leq$ RHS. Hence LHS equals RHS.

Now we proof the lemma.

Proof of Lemma 1. We will prove by induction on the lengths of $\mathbf{s}$ and $\mathbf{t}$, the base case when $\#\mathbf{s}, \#\mathbf{t} \le 2$ can be readily checked.

Assume that the statement holds for every $\mathbf{s}$ and $\mathbf{t}$ such that $\#\mathbf{s} \le M - 1$ and $\#\mathbf{t} \le N$, it remains to show that the statement is true for any $\mathbf{u}$ and $\mathbf{t}$ such that $\#\mathbf{u} = M$ and $\#\mathbf{v} = N$.

Let $\mathbf{u}' \in \mathcal{S}(\mathbf{u})$, $\mathbf{v}' \in \mathcal{S}(\mathbf{v})$, $\#\mathbf{u}' = \#\mathbf{v}' > \#\mathbf{u} + \#\mathbf{v} - 1$. Then $\mathbf{u}'$ can be written as $\mathbf{u}_{\mathbf{h}'}\mathbf{u}_{\mathbf{t}'}$ where every element of $\mathbf{u}_{\mathbf{h}'}$ equals u_1 and the first element of $\mathbf{u}_{\mathbf{t}'}$ equals u_2 . Since $\mathbf{v}'$ is a stretch

of $\mathbf{v}$, $\mathbf{v}'\langle\#\mathbf{u}_{\mathbf{h}'}\rangle$ must be some element in $\mathbf{v}$, say v_k. There are two possibilities of $\mathbf{v}'\langle\#\mathbf{u}_{\mathbf{h}'}+1\rangle$. For the case $\mathbf{v}'\langle\#\mathbf{u}_{\mathbf{h}'}+1\rangle = v_k$, $\mathbf{u}'$ and $\mathbf{v}'$ are aligned as,

$$\mathbf{u}' = [\underbrace{u_1,\ldots,u_1}_{\mathbf{u}_{\mathbf{h}'}},\underbrace{u_2,\ldots,u_M}_{\mathbf{u}_{\mathbf{t}'}}]$$
$$\mathbf{v}' = [\underbrace{v_1,\ldots,v_k}_{\mathbf{v}_{\mathbf{h}'}},\underbrace{v_k,\ldots,v_N}_{\mathbf{v}_{\mathbf{t}'}}].$$

Define $\mathbf{v}_{\mathbf{h}}$ as the the sequence $[v_1,v_2,\ldots,v_k]$ and $\mathbf{v}_t$ as the tail of $\mathbf{v}$ from the k-th element onwards. One can check that $\mathbf{u}'_{\mathbf{t}} \in \mathcal{S}(\mathbf{u}_{\sim})$ and $\mathbf{v}'_{\mathbf{t}} \in \mathcal{S}(\mathbf{v}_{\mathbf{t}})$. Since $\#\mathbf{u}_{\sim} \le M-1$ and $\#\mathbf{v}_{\mathbf{t}} \le N$, by the assumption, there are $\mathbf{u}''_{\mathbf{t}}$ and $\mathbf{v}''_{\mathbf{t}}$ such that

$$\begin{aligned}\#\mathbf{u}''_{\mathbf{t}} = \#\mathbf{v}''_{\mathbf{t}} &= \#\mathbf{u}_{\sim} + \#\mathbf{v}_{\mathbf{t}} - 1\\ &= (\#\mathbf{u}-1)+(\#\mathbf{v}-k+1)-1\\ &= \#\mathbf{u}+\#\mathbf{v}-k-1,\end{aligned}$$

and $\|\mathbf{u}''_{\mathbf{t}} - \mathbf{v}''_{\mathbf{t}}\|_p \le \|\mathbf{u}'_{\mathbf{t}} - \mathbf{v}'_{\mathbf{t}}\|_p$. Write $\mathbf{u}'' = [u_1]^k\mathbf{u}''_{\mathbf{t}}$ and $\mathbf{v}'' = \mathbf{v}_{\mathbf{h}}\mathbf{v}''_{\mathbf{t}}$. It is easy to see that $\mathbf{u}'' \in \mathcal{S}(u)$ and $\mathbf{v}'' \in \mathcal{S}(v)$.

Furthermore, $\mathbf{u}'' = \mathbf{v}'' = \#\mathbf{u}+\#\mathbf{v}-1$ and

$$\begin{aligned}\|\mathbf{u}''-\mathbf{v}''\|_p^p &= \|[u_1]^k - \mathbf{v}_{\mathbf{h}}\|_p^p + \|\mathbf{u}''_{\mathbf{t}} - \mathbf{v}''_{\mathbf{t}}\|_p^p\\ &\le \|[u_1]^{\#\mathbf{u}'_{\mathbf{h}}} - \mathbf{v}'_{\mathbf{h}}\|^p + \|\mathbf{u}'_{\mathbf{t}} - \mathbf{v}'_{\mathbf{t}}\|_p^p\\ &= \|\mathbf{u}'-\mathbf{v}'\|_p^p,\end{aligned}$$

for $p \in [1,\infty)$, and,

$$\begin{aligned}\|\mathbf{u}''-\mathbf{v}''\|_\infty &= \max\left(\|[u_1]^k - \mathbf{v}_{\mathbf{h}}\|_\infty, \|\mathbf{u}''_{\mathbf{t}} - \mathbf{v}''_{\mathbf{t}}\|_\infty\right)\\ &\le \max\left(\|[u_1]^{\#\mathbf{u}'_{\mathbf{h}}} - \mathbf{v}'_{\mathbf{h}}\|_\infty, \|\mathbf{u}'_{\mathbf{t}} - \mathbf{v}'_{\mathbf{t}}\|_\infty\right)\\ &= \|\mathbf{u}'-\mathbf{v}'\|_\infty .\end{aligned}$$

So $\|\mathbf{u}''-\mathbf{v}''\|_p \le \|\mathbf{u}'-\mathbf{v}'\|_p$.

For the case $\mathbf{v}'\langle\#\mathbf{u}_{\mathbf{h}'}+1\rangle = v_{k+1}$, we can proceed through a similar argument and have $\mathbf{u}''' \in \mathcal{S}(u)$ and $\mathbf{v}''' \in \mathcal{S}(v)$ whose lengths are equal to $\#\mathbf{u}+\#\mathbf{v}-2$ and $\|\mathbf{u}'''-\mathbf{v}'''\|_p \le \|\mathbf{u}'-\mathbf{v}'\|_p$. ■

As a by-product of the previous discussion we have an alternative characterization of the DTW, for $p \in [1,\infty)$,

$$\mathrm{DTW}(\mathbf{x},\mathbf{y}) = \inf_{\mu,\nu\in\mathcal{S}} \|\mu\mathbf{x} - \nu\mathbf{y}\|_p . \tag{8}$$

Where $\|\mathbf{x}-\mathbf{y}\|_p$ is implicitly defined to be ∞ when $\#\mathbf{x} \ne \#\mathbf{y}$.

Indeed, DTW is the condensation of ℓ^p metric wrt. the set of stretch operations $\mathcal{S}$, but since $\mathcal{S}$ does not preserve the ℓ^p metric for $p \in [1,\infty)$ Theorem 3 will not guarantee that DTW is subadditive and it is actually not subadditive as we have seen in Section 2.

The next example is the condensations of ℓ^p metrics wrt. *gap* insertions.

doi:10.4186/ej.2009.13.2.19

Definition 7. *Let* γ *be a real number and* W *be the set of sequences with constant tail* γ*, i.e. the sequences of the form* $(x_1,\ldots,x_l,\gamma,\gamma,\ldots)$*. Define the map* $\iota_0 : W \to W$ *by,*

$$\iota_0\mathbf{x} = (\gamma, x_1, x_2, \ldots).$$

For $k \in \mathbb{N}$ *define* $\iota_k : W \to W$ *by,*

$$\iota_k\mathbf{x} = (x_1,\ldots,x_k,\gamma,x_{k+1},\ldots).$$

Precisely, $\iota_k\mathbf{x}\langle k+1\rangle = \gamma$*,* $\iota_k\mathbf{x}\langle i\rangle = \mathbf{x}\langle i\rangle$ *for* $1 \le i \le k$*, and* $\iota_k\mathbf{x}\langle j\rangle = \mathbf{x}\langle j-1\rangle$ *for* $j \ge k+2$*.*

Let $\mathcal{I}_\gamma$ be the set of all finite compositions of operations in $\{\iota_k\}_{k\ge 0} \cup \{\mathbb{I}\}$. We sometimes write $\mathcal{I}_\gamma$ as $\mathcal{I}$ when there is no need to specify the value γ.

W can be thought of as either a set of infinite sequences with constant tails as in the definition above or as a set of finite sequences and let the distance between sequences of different lengths be ∞ and use the ℓ^p metric as the distance if the sequences are of the same length.

An illustration of gap insertions is in Figure 6. One can check that $\mathcal{I}$ is complete and it preserves ℓ^p metrics. Hence, the condensation of ℓ^p metrics wrt. $\mathcal{I}$,

$$\delta_2^p(\mathbf{x},\mathbf{y}) = \inf_{\iota,\kappa\in\mathcal{I}} \| \iota\mathbf{x} - \kappa\mathbf{y} \|_p, \tag{9}$$

is a pseudometric. It can be computed by dynamic programming using the relation below,

$$\delta_2^p(\mathbf{s},\mathbf{t})^p = \min\begin{cases} |\, s_1 - t_1 \,|^p + \delta_2^p(\mathbf{s}_\sim,\mathbf{t}_\sim)^p, \\ |\, \gamma - t_1 \,|^p + \delta_2^p(\mathbf{s},\mathbf{t}_\sim)^p, \\ |\, s_1 - \gamma \,|^p + \delta_2^p(\mathbf{s}_\sim,\mathbf{t})^p. \end{cases} \tag{10}$$

Where $\delta_2^p([],\mathbf{s})^p = \delta_2^p(\mathbf{s},[])^p = \sum_{i=1}^{l} |\, \gamma - s_i \,|^p$ for the base cases. $\delta_2(\mathbf{s},\mathbf{t})$ can be computed in $O(\#\mathbf{s}\#\mathbf{t})$.

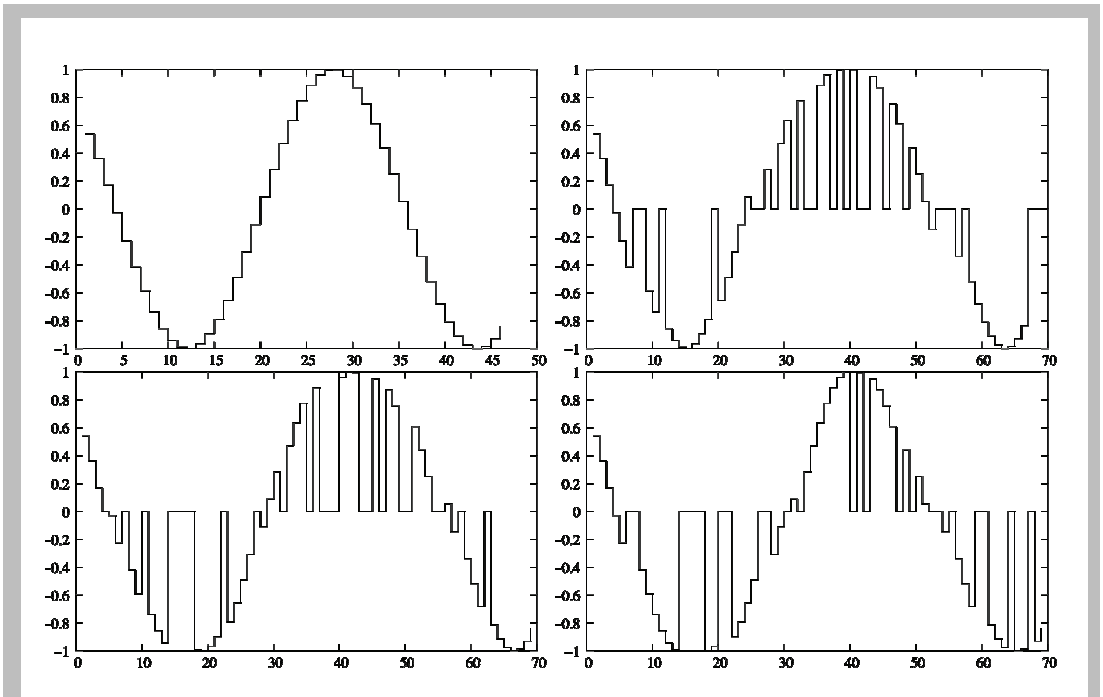

Figure 6
A visualization of a time series and its possible results after insertion operations with functions in the class $\mathcal{I}$. The original time series is the top-left. The rest are some of its possible results after gap insertions. The gap value is 0.

The fact that the recurrence relation solves the distance defined in Equation (9) follows from an argument similar to that used to explain the case of δ_1. However, note that the recurrence relation computes $\inf_{\iota,\kappa\in\mathcal{I}} \| \iota\mathbf{x} - \kappa\mathbf{y} \|_p^p$. Since a^p is strictly increasing when a is nonnegative, it is possible to minimize $\| \iota\mathbf{x} - \kappa\mathbf{y} \|_p^p$ instead of $\| \iota\mathbf{x} - \kappa\mathbf{y} \|_p$. Indeed one can check that

$$\delta_2^p(\mathbf{x},\mathbf{y}) = \inf_{\iota,\kappa\in\mathcal{I}} \| \iota\mathbf{x} - \kappa\mathbf{y} \|_p = \left(\inf_{\iota,\kappa\in\mathcal{I}} \| \iota\mathbf{x} - \kappa\mathbf{y} \|_p^p \right)^{\frac{1}{p}}.$$

Mark that the value of p can be any real number, but δ_2 will be subadditive if $p \in [1,\infty)$ because then the ℓ^p distance will be subadditive and δ_2 will be a condensation of a metric. For the particular case of $p = 1$, this is the distance function known as ERP [16] that we mentioned in Subsection 2.9.

Consider the case when $p = 2$ and $\gamma = 0$, in computing the distance δ_2^2 between $\mathbf{x}$ and $\mathbf{y}$. The quantity $\| \iota\mathbf{x} - \kappa\mathbf{y} \|_2$ is minimized over all possible insertions ι and κ in $\mathcal{I}_0$. Recalling that the square function is strictly increasing when the domain is positive, it is possible to minimize $\| \iota\mathbf{x} - \kappa\mathbf{y} \|_2^2$ instead, and by the polarization identity,

$$\| \iota\mathbf{x} - \kappa\mathbf{y} \|_2^2 = \| \iota\mathbf{x} \|_2^2 + \| \kappa\mathbf{y} \|_2^2 - 2\iota\mathbf{x} \cdot \kappa\mathbf{y}.$$

Since $\gamma = 0$ gap insertions keep the norm of time series unchanged, i.e. $\iota\mathbf{x}$ for every ι in $\mathcal{I}_0$ and $\mathbf{x}$ in W. Therefore we have

$$\| \iota\mathbf{x} - \kappa\mathbf{y} \|_2^2 = \| \mathbf{x} \|_2^2 + \| \mathbf{y} \|_2^2 - 2\iota\mathbf{x} \cdot \kappa\mathbf{y}.$$

So minimizing the quantity above over all ι and κ in $\mathcal{I}_0$ is the same as maximizing the dot product $\iota\mathbf{x} \cdot \kappa\mathbf{y}$ in the right hand side of the above equation over all ι and κ in $\mathcal{I}_0$.

To summarize, for the case when $p = 2$ and $\gamma = 0$ the distance δ_2^2 between $\mathbf{x}$ and $\mathbf{y}$ can be viewed as the ℓ^2 distance between the *inserted* time series $\iota\mathbf{x}$ and $\kappa\mathbf{y}$ derived from $\mathbf{x}$ and $\mathbf{y}$ such that their similarity as measured by their dot product $\iota\mathbf{x} \cdot \kappa\mathbf{y}$ is maximized.
Inspired by the above discussion, we propose another subadditive condensation based on the idea of maximizing similarity. We first introduce the distance function

$$\angle(\mathbf{x},\mathbf{y}) = \arccos\left(\frac{\mathbf{x} \cdot \mathbf{y}}{\| \mathbf{x} \|_2 \| \mathbf{y} \|_2} \right),$$

whose geometric interpretation is the measure of the angle between two vectors. One can check that $\angle$ is a pseudometric.

Now we condense the distance function wrt. the *insertions* $\mathcal{I}_0$ defined above with the gap value $\gamma = 0$, giving the condensation

$$\begin{aligned} \delta_3(\mathbf{x},\mathbf{y}) &= \inf_{\iota,\kappa\in\mathcal{I}_0} \angle(\iota\mathbf{x},\kappa\mathbf{y}) \\ &= \inf_{\iota,\kappa\in\mathcal{I}_0} \arccos\left(\frac{\iota\mathbf{x} \cdot \kappa\mathbf{y}}{\| \iota\mathbf{x} \|_2 \| \kappa\mathbf{y} \|_2} \right) \\ &= \inf_{\iota,\kappa\in\mathcal{I}_0} \arccos\left(\frac{\iota\mathbf{x} \cdot \kappa\mathbf{y}}{\| \mathbf{x} \|_2 \| \mathbf{y} \|_2} \right). \end{aligned}$$

doi:10.4186/ej.2009.13.2.19

The third equality follows from the fact that insertions of zero gaps preserves the norms. When the gap value γ is zero, it can be checked that the insertions of gaps preserve the distance $\angle$. Hence it follows that δ_3 is a pseudometric.

For every real number $\alpha > 0$ and every pair of time series $\mathbf{s}$ and $\mathbf{s}'$ such that $\mathbf{s}' = \alpha\mathbf{s} = [\alpha s_1, \ldots, \alpha s_l]$ we have $\delta_3(\mathbf{s}, \mathbf{s}') = 0$. To justify this intuitively, if we have a time series $\mathbf{s}$ and another one with the same *shape* but of different scales, then they are not different when measured with the distance δ_3.

Next we briefly discuss a way to compute δ_3. First note that the arccos function is strictly decreasing, this fact can be used to show that

$$\delta_3(\mathbf{x}, \mathbf{y}) = \cos\left(\sup_{\iota,\kappa \in \mathcal{I}_0} \frac{\iota\mathbf{x} \cdot \kappa\mathbf{y}}{\| \mathbf{x} \|_2 \| \mathbf{y} \|_2} \right).$$

So one can maximize the quantity $\iota\mathbf{x} \cdot \kappa\mathbf{y}$ over all possible ι and κ in $\mathcal{I}_0$ instead. Writing the quantity $\sup_{\iota,\kappa \in \mathcal{I}_0} \iota\mathbf{x} \cdot \kappa\mathbf{y}$ as $\delta_{iii}(\mathbf{x}, \mathbf{y})$, we can then compute the distance δ_3 by

$$\delta_3(\mathbf{x}, \mathbf{y}) = \cos\left(\delta_{iii}(\mathbf{x}, \mathbf{y}) / \| \mathbf{x} \|_2 \| \mathbf{y} \|_2 \right).$$

δ_{iii} can be computed using the recurrence relation

$$\delta_{iii}(\mathbf{s}, \mathbf{t}) = \max \begin{cases} s_1 t_1 + \delta_{iii}(\mathbf{s}_{\sim}, \mathbf{t}_{\sim}), \\ \delta_{iii}(\mathbf{s}, \mathbf{t}_{\sim}), \\ \delta_{iii}(\mathbf{s}_{\sim}, \mathbf{t}). \end{cases}$$

Where $\delta_{iii}([], \mathbf{s}) = \delta_{iii}(\mathbf{s}, []) = 0$ for the base cases.

Again, $\delta_{iii}(\mathbf{s}, \mathbf{t})$ can be computed in $O(\#\mathbf{s}\#\mathbf{t})$ time. Since δ_3 can be computed from δ_{iii} in constant time, δ_3 can also be computed in $O(\#\mathbf{s}\#\mathbf{t})$ time.

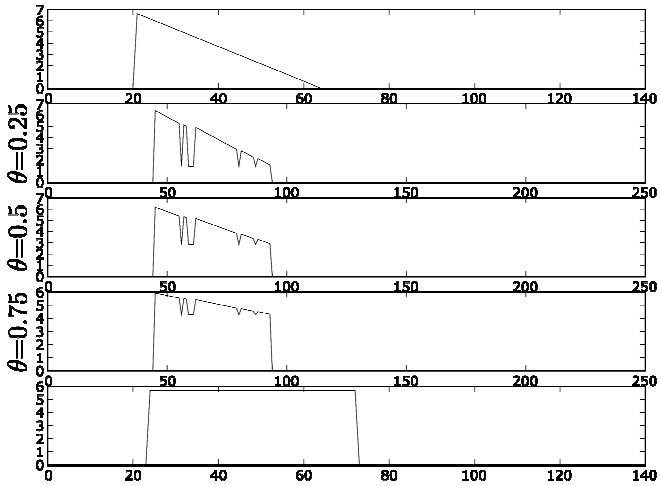

Figure 7
The interpolation between two time series wrt. δ_2^1. The value γ is set to zero. When θ is closer to 1 the interpolated time series is closer to the bottom time series when measured with δ_2^1.

3.1.2 Interpolation of Time Series

If a condensation is well defined in the form

$$\underset{d,\mathcal{M}}{\mathrm{D}}(x,y) = \min_{\mu,\nu\in\mathcal{M}} \| \mu x - \nu y \|,$$

or, to put it differently, if the pair of morph operations yielding minimal distance always exists and the base distance is a norm metric, then we can do interpolation of objects in a certain way. δ_1 and δ_2 are examples of this type of condensation.

Proposition 1. *Let* $\mathbf{a},\mathbf{b}$ *be vectors in a vector space* V *with a norm* $\|\cdot\|$. *Let* $\mathcal{M}$ *be a complete set of morph operations on* V *preserving the* $\|\cdot\|$ *-metric. If* $\Delta(\mathbf{x},\mathbf{y}) = \min_{\mu,\nu\in\mathcal{M}} \| \mu x - \nu y \|$ *is well defined, then for any* $\theta \in [0,1]$ *there is* $\mathbf{c} \in V$ *satisfying* $\Delta(\mathbf{a},\mathbf{c}) = \theta\Delta(\mathbf{a},\mathbf{b})$ *and* $\Delta(\mathbf{b},\mathbf{c}) = (1-\theta)\Delta(\mathbf{a},\mathbf{b})$.

Proof. By assumption there are $\mu_0,\nu_0 \in \mathcal{M}$ yielding $\| \mu_0\mathbf{a} - \nu_0\mathbf{b} \| = \Delta(\mathbf{a},\mathbf{b})$. Write $\mathbf{a}' = \mu_0\mathbf{a}$ and $\mathbf{b}' = \nu_0\mathbf{b}$. Let $\mathbf{c} = (1-\theta)\mathbf{a}' + \theta\mathbf{b}'$, then

$$\begin{aligned} \Delta(\mathbf{a},\mathbf{c}) &\leq \| \mathbf{a}' - \mathbf{c} \| = \theta \| \mathbf{a}' - \mathbf{b}' \| = \theta\Delta(\mathbf{a},\mathbf{b}), \\ \Delta(\mathbf{c},\mathbf{b}) &\leq \| \mathbf{c} - \mathbf{b}' \| = (1-\theta) \| \mathbf{b}' - \mathbf{a}' \| = (1-\theta)\Delta(\mathbf{a},\mathbf{b}). \end{aligned}$$

Since Δ is subadditive (by Theorem 3), and by the two inequalities above,

$$\begin{aligned} \Delta(\mathbf{a},\mathbf{b}) &\leq \Delta(\mathbf{a},\mathbf{c}) + \Delta(\mathbf{c},\mathbf{b}) \\ &\leq \theta\Delta(\mathbf{a},\mathbf{b}) + (1-\theta)\Delta(\mathbf{a},\mathbf{b}) = \Delta(\mathbf{a},\mathbf{b}). \end{aligned}$$

This implies that $\Delta(\mathbf{a},\mathbf{c}) \geq \theta\Delta(\mathbf{a},\mathbf{b})$ and $\Delta(\mathbf{b},\mathbf{c}) \geq (1-\theta)\Delta(\mathbf{a},\mathbf{b})$. Together with the first two inequalities, we conclude that $\Delta(\mathbf{a},\mathbf{c}) = \theta\Delta(\mathbf{a},\mathbf{b})$ and $\Delta(\mathbf{b},\mathbf{c}) = (1-\theta)\Delta(\mathbf{a},\mathbf{b})$. ■

The proposition says that a way to interpolate between two time series **s** and **t** wrt. $\Delta_{d,\mathcal{M}}$ is by doing linear interpolation between the closest pair of time series among all possible pairs such that one of the pair can be morphed from **s** and the other can be morphed from **t**. Figure 7 shows an example of interpolations between two time series wrt. the distance δ_2.

3.2 Shortcut Distance

Lemma 2. *Given a set* Ω *and a nonnegative function* $d : \Omega\times\Omega \to [0,\infty]$ *such that* $d(x,x) = 0$ *for every* x *in* Ω. *The function* $\Xi_d : \Omega\times\Omega \to [0,\infty]$ *defined by,*

$$\Xi_d(x,y) = \inf\left\{ \sum_{i=1}^{N} d(x_{i-1},x_i) \mid x_0,\ldots,x_n \in \Omega, N \in \mathbb{N}, x_0 = x, x_n = y \right\}$$

is subadditive. Ξ_d *is called the* shortcut *of* d.

Proof. Let p,q,r be any points in Ω. For a fixed $\varepsilon > 0$, by definition there are $\{p = p_0,\ldots,p_M = q = q_0,\ldots,r = q_N\} \subseteq \Omega$ such that,

$$\Xi_d(p,q) + \Xi_d(q,r) + \varepsilon \geq \sum_{i=1}^{M} d(p_{i-1},p_i) + \sum_{j=1}^{N} d(q_{j-1},q_j).$$

doi:10.4186/ej.2009.13.2.19

Rename the points $q_1,\ldots,q_N$ to $p_{M+1},\ldots,p_{M+N}$ respectively. Noting that $d(p_M,q_0)=0$, we can write the above inequality as,

$$\Xi_d(p,q)+\Xi_d(q,r)+\varepsilon \geq \sum_{i=1}^{M+N} d(p_{i-1},p_i) \geq \Xi_d(p,r).$$

The second inequality holds by definition. Since this is true for arbitrary $\varepsilon > 0$ we conclude that Ξ_d is subadditive. ■

One can check that if d is symmetric, then Ξ_d is also symmetric. By the lemma, we can always construct a pseudometric from a symmetric distance function by defining its shortcut. Even if the distance function d we have is not symmetric, we can create a symmetric distance function first and use the new symmetric distance. For example, the functions $(x,y)\mapsto \frac{1}{2}d(x,y)+\frac{1}{2}d(y,x)$ and $(x,y)\mapsto \max\{d(x,y),d(y,x)\}$ are always symmetric, and they will be subadditive if d is subadditive. In fact, however, symmetry is not a vital property because one can do nearest neighbor anyway by explicitly specifying the direction when comparing distances.

Note that a subadditive distance function is always a shortcut of some distance function since it is always a shortcut of itself. Together with the lemma this suggests a vague intuition that every metric measures the minimum cost of finite gradual changes from one object to another. In a sense, the shortcut is the length of the *shortest path* between objects, and for subadditive distances, the shortest path is always the *direct* one.

The question of what is the shortcut of DTW is an open question; once the shortcut of DTW is known, the way to efficiently compute it and its classification performance are the next to be enquired.

3.2.1 Examples

By inspecting the expression in Equation (10) of δ_2 (ERP) with the assumption that γ in the equation is zero and the base metric is the ℓ^1 metric, it can be shown that $\delta_2^1(\mathbf{x},\mathbf{y})$ is the minimum cost of the sequence of the following transformations leading $\mathbf{x}$ to $\mathbf{y}$,

- delete x_i from $\mathbf{x}$, resulting a shorter time series, costs $|x_i|$,
- insert a number r into $\mathbf{x}$, resulting a longer time series, costs $|r|$,
- change the value of x_i to v, costs $|x_i - v|$.

The right hand side of Equation (10) chooses the minimum among three choices. We can think of the first choice as the cost we have to pay in order to transform $\mathbf{s}$ to $\mathbf{t}$ by changing the value s_1 of the first element of $\mathbf{s}$ to match t_1 and do the best we can for the rest. The second choice is the lowest cost we need to pay if we insert t_1 at the head of $\mathbf{s}$ first. The third choice is for the case of deleting the first element of $\mathbf{s}$ first.

Observe that the minimum cost of transformations from $\mathbf{x}$ to $\mathbf{y}$ is equal to the cost of transformations from $\mathbf{y}$ to $\mathbf{x}$. This is because there is a sequence of transformations from $\mathbf{y}$ to $\mathbf{x}$ with cost c for each sequence that transforms $\mathbf{x}$ to $\mathbf{y}$ whose cost is also c. Suppose we have a sequence of transformations making $\mathbf{x}$ become $\mathbf{y}$, we can reverse the order of that sequence and substitute each deletion with an appropriate insertion and vice versa as well as reversing the changes of values. The new sequence of transformations will have equal cost and change $\mathbf{y}$ to $\mathbf{x}$.

For $\pi > 0$, the minimum cost of the sequence of the following morphs leading $\mathbf{x}$ to $\mathbf{y}$ is another pseudometric

- delete x_i from $\mathbf{x}$, costs $|x_i| + \pi$,
- insert a number r into $\mathbf{x}$, costs $|r| + \pi$,
- change the value of x_i to v, costs $|x_i - v|$.

The distance above is more general than the ERP distance because it reduces to ERP when π is zero. The value π can be thought of as the penalty needed to be paid for each insertion, and deletion.

The minimum cost can be found by dynamic programming using the relation below.

$$\delta_4(\mathbf{s},\mathbf{t}) = \min\begin{cases} |s_1 - t_1| + \delta_4(\mathbf{s}_{\sim},\mathbf{t}_{\sim}), \\ \pi + |t_1| + \delta_4(\mathbf{s},\mathbf{t}_{\sim}), \\ \pi + |s_1| + \delta_4(\mathbf{s}_{\sim},\mathbf{t}). \end{cases}$$

Where $\delta_4([],\mathbf{s}) = \delta_4(\mathbf{s},[]) = l\pi + \sum_{i=1}^{l} |s_i|$ for the initial cases. $\delta_4(\mathbf{s},\mathbf{t})$ can be computed in $O(\#\mathbf{s}\#\mathbf{t})$ time.

In fact we can have a more general form

$$\delta_4^p(\mathbf{s},\mathbf{t})^p = \min\begin{cases} |s_1 - t_1|^p + \delta_4(\mathbf{s}_{\sim},\mathbf{t}_{\sim})^p, \\ \pi + |t_1|^p + \delta_4(\mathbf{s},\mathbf{t}_{\sim})^p, \\ \pi + |s_1|^p + \delta_4(\mathbf{s}_{\sim},\mathbf{t})^p. \end{cases}$$

Where $\delta_4^p([],\mathbf{s}) = \delta_4^p(\mathbf{s},[]) = \left(l\pi + \sum_{i=1}^{l} |s_i|^p\right)^{1/p}$ for the initial cases.

We can show that $\delta_4^p(\mathbf{x},\mathbf{y})$ computes the p-th root of the minimum cost of the sequence of the following transformations leading $\mathbf{x}$ to $\mathbf{y}$,

- delete x_i from $\mathbf{x}$, costs $|x_i|^p + \pi$,
- insert a number r into $\mathbf{x}$, costs $|r|^p + \pi$,
- change the value of x_i to v, costs $|x_i - v|^p$.

Using the fact that $(a+b)^{1/p} \le a^{1/p} + b^{1/p}$ for $a,b \ge 0$ and $p \ge 1$, one can check that the distance δ_4^p is subadditive for every $p \ge 1$.

As a final remark, we note again that the argument that the set of all time series with all rational values is a countable dense subset will suffice to establish that all of the pseudometric spaces of time series in this section are separable. Consequently, Theorem 1 about asymptotic properties applies to k-NN in these pseudometrics.

IV. Numerical Results

We compare the performance of our distance measures with the Euclidean distance and DTW using the datasets from the UCR time series classification/clustering homepage [26]. The experimental setting is done as in the UCR dataset homepage. The 1-NN algorithm was used to perform classification tasks. Each of the datasets is already split as the training set and the testing set. Numbers shown in each column of distance measure are the error rates as classified by the 1-NN algorithm using that distance function as the dissimilarity measure. The results are shown in Table 1.

doi:10.4186/ej.2009.13.2.19

Every time series in each dataset was already pre-scaled to the same length if needed so that the ℓ^1 distance and the Euclidean distance can be computed between every pair of time series in each dataset.

The distance δ_1 was incompetent because its base distance, the ℓ^∞ metric is usually very sensitive to little variations or noise in time series. Other than δ_1, our proposed subadditive distances did relatively well. The results also showed that the value of π and *p* had influence on the classification accuracy of δ_4.

Nonetheless, it is interesting to note that the *trace* dataset can be classified with zero error rate by DTW and δ_1. The identical property shared by the two distances is that they are condensations whose morphs are the stretch operations $\mathcal{S}$ (cf.Equation (5) and Equation (8)). While the condensations δ_{2-4}, whose morphs are the gap insertions $\mathcal{I}$, gave relatively poor performance. A plausible explanation of the phenomenon is that the stretches are more suitable as morph operations than gap insertions for this particular dataset. It seemed, though, that some datasets have no preference of either the stretches or the insertions over the other, the *two patterns* dataset can be classified by 1-NN with zero error rate with each condensation regardless of the morphs involved in the condensation.

It is noticeable that the results of the distance δ_2^2 and the distance δ_3 are the same or almost the same in many datasets. This is due to the fact that if the ℓ^2 norm of every time series in a dataset is the same, then the nearest neighbors of a time series as measured by δ_2^2 will also be the nearest neighbors as measured by δ_3 and vice versa. Those datasets whose time series have small variation among their ℓ^2 norms yielded the same error rates for δ_2^2 and δ_3. For those datasets whose time series have large variation among their ℓ^2 norms the results were different. The time series in the coffee dataset have very large variation of the ℓ^2 norms and the distance δ_3 performed significantly better than other distances including δ_2^2.

V. Conclusion

Pseudometrics or subadditive distances are both theoretically and practically appealing. Asymptotically, pseudometric based *k* -NN has the chance to have error no more than twice that of the Bayes classifier, and several implementations can be used to accelerate pseudometric based *k* -NN. We provide two frameworks for designing subadditive distance measures for time series, namely the *condensation* framework and the *shortcut distance* framework.

Condensation of a distance is a new distance based on an existing one, the base distance. In the condensation framework one can get a subadditive distance by choosing an appropriate set of morph operations wrt. the base distance. In order to have a subadditive condensation of a distance *d* wrt. a set $\mathcal{M}$ of morph operations, one need to check that the three following conditions are valid,

1. the base distance *d* is subadditive,
2. the morphs $\mathcal{M}$ is complete,
3. $\mathcal{M}$ contracts *d*.

DTW can be regarded as a condensation, but since the morph operations involved are not appropriate, DTW is not subadditive. An existing pseudometric called ERP and its generalization can be constructed in this framework. For norm metric based condensations, one can do interpolation between time series in a way similar to linear interpolation.

By two tools we proposed, Theorem 3 and Lemma 1, we designed our example pseudometrics in a somewhat modular fashion. The condensations were developed and we just checked the three conditions of Theorem 3 without worrying that some expression involving infinity may trouble the possibility of real implementation. By Lemma 1, and maybe its modifications, for some well-behaved morphs we can reduce computations involving infinity to finite ones, and fortunately all the morphs in our example condensations are such well-behaved.

The second framework based on shortcut distances is more general. Any shortcut of a distance is always subadditive. Moreover, any subadditive distance is an edit distance in disguise. A concrete definition of the shortcut of DTW and its algorithm are currently unknown. A more general form of an existing distance called ERP can also be constructed in the second framework. We fine tuned it by adding a penalty value to prevent too much *morphs* to match another time series. The fine tuned distance has potential to yield better classification results. All of the proposed distances can be computed in $O(mn)$ time like DTW. Numerical results showed that they are useful alternatives to DTW.

doi:10.4186/ej.2009.13.2.19

Table 1 Numbers whose value is minimal in its row are typeset bold. The second column shows the standard deviations of the ℓ^2 norms of the time series samples in the training set of each dataset.

dataset	$\sigma(\|\cdot\|_2)$	ℓ^1	ℓ^2	DTW^1	DTW^2	δ_1	δ_2^1	δ_2^2	δ_3	δ_4^1 $\pi=0.5$	δ_4^1 $\pi=1.0$	δ_4^2 $\pi=0.5$	δ_4^2 $\pi=1.0$
50words	2.512e-08	0.3319	0.3692	0.2835	0.3099	0.4264	0.2813	0.3231	0.3231	0.2352	0.2242	**0.2110**	0.2462
adiac	2.449e-08	0.4015	0.3887	0.4118	0.3964	0.4220	0.3785	**0.3708**	**0.3708**	0.3785	0.3939	0.3862	0.3887
beef	1.037e+00	0.5000	0.4667	0.5000	0.5000	0.4333	0.5000	0.5000	**0.2333**	0.5000	0.5000	0.4667	0.4667
cbf	2.637e-08	0.1111	0.1478	**0.0000**	0.0033	0.0289	0.0022	0.0022	0.0022	0.0100	0.0322	0.0078	0.0156
coffee	4.643e+01	0.2500	0.2500	0.1786	0.1786	0.3214	0.2500	0.2500	**0.0357**	0.2500	0.2500	0.2500	0.2500
ecg200	6.052e-03	0.1100	0.1200	0.2000	0.2300	0.2500	0.1500	0.1800	0.1800	0.1600	**0.1000**	0.1400	0.1300
faceall	2.465e-08	0.2787	0.2864	0.2284	0.1923	0.2379	0.2024	0.2030	0.2036	0.1941	0.1947	**0.1882**	0.1935
facefour	3.670e-08	0.1591	0.2159	0.1591	0.1705	0.4205	0.1023	0.1705	0.1705	0.0568	**0.0455**	0.1364	0.1250
fish	1.068e-01	0.2057	0.2171	0.1429	0.1657	0.3029	**0.1200**	0.1314	0.1314	0.1371	0.1600	0.2000	0.2114
gun point	2.311e-08	0.0467	0.0867	0.1200	0.0933	0.2133	**0.0400**	**0.0400**	**0.0400**	0.0467	0.0467	0.0800	0.0867
lighting2	1.350e-07	0.1803	0.2459	0.1967	0.1311	0.2787	0.1475	**0.1148**	**0.1148**	**0.1148**	**0.1148**	0.1311	**0.1148**
lighting7	1.210e-07	0.2877	0.4247	**0.2329**	0.2740	0.5205	0.3014	0.3288	0.3288	0.2740	0.2603	0.2466	0.2603
oliveoil	3.710e-02	0.1667	**0.1333**	**0.1333**	**0.1333**	0.2000	0.1667	**0.1333**	0.1667	0.1667	0.1667	**0.1333**	**0.1333**
osuleaf	5.997e-03	0.4504	0.4835	0.3678	0.4091	0.4876	0.3967	0.4132	0.4132	**0.3512**	0.3636	**0.3512**	0.3719
swedishleaf	2.513e-06	0.2112	0.2112	0.2096	0.2080	0.2464	0.1200	0.1232	0.1232	**0.1120**	0.1264	0.1184	0.1408
synthetic control	2.614e-08	0.1200	0.1200	0.0133	**0.0067**	0.0467	0.0333	0.0300	0.0300	0.0267	0.0367	0.0300	0.0233
trace	2.585e-08	0.2400	0.2400	0.0100	**0.0000**	**0.0000**	0.1700	0.1600	0.1600	0.1800	0.2000	0.1700	0.1700
two patterns	1.392e-07	0.0387	0.0932	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
wafer	1.274e-07	0.0047	0.0045	0.0161	0.0201	0.0339	0.0088	0.0092	0.0092	0.0049	**0.0028**	0.0049	**0.0028**
yoga	2.620e-08	0.1710	0.1697	0.1613	0.1637	0.1757	0.1470	0.1623	0.1623	**0.1307**	0.1363	0.1487	0.1583

REFERENCES

[1] E. Fix and J. L. Hodges "Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties," USAF School of Aviation Medicine, Randolf Field, Texas, Technical Report Project 21-49-004, Report Number 4, 1951.

[2] E. Fix and J. L. Hodges, Jr., "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties," *International Statistical Review / Revue Internationale de Statistique*, vol. 57, pp. 238-247, 1989.

[3] B. L. Welch, "(ii) Note on Discriminant Functions," *Biometrika*, vol. 31, pp. 218, 1939.

[4] T. Cover and P. Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on*, vol. 13, pp. 21-27, 1967.

[5] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in Neural Information Processing Systems 17*: MIT Press, pp. 513-520, 2004.

[6] K. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *Advances in Neural Information Processing Systems*, vol. 18, pp. 1473-1480, 2006.

[7] W. Zhang, X. Xue, Z. Sun, Y.-F. Guo, and H. Lu, "Optimal dimensionality of metric space for classification," in *Proceedings of the 24th international conference on Machine learning*. Corvalis, Oregon: ACM, pp. 1135-1142, 2007.

[8] R. Chatpatanasiri, T. Korsrilabutr, P. Tangchanachaianan, and B. Kijsirikul, "On kerneliz ation of supervised mahalanobis distance learners," 2008.

[9] J. Barros, J. French, W. Martin, P. Kelly, and M. Cannon, "Using the triangle inequality to reduce the number of comparisons required for similarity-based retrieval," in *Proceedings of SPIE/IS&T Conference on Storage and Retrieval for Image and Video Databases IV*: SPIE, pp. 392-403, 1996.

[10] N. Roussopoulos, S. Kelley, and F. d. r. Vincent, "Nearest neighbor queries," in *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*. San Jose, California, United States: ACM, pp. 71-79, 1995.

[11] P. Ciaccia, M. Patella, and P. Zezula, "M-tree: An efficient access Method for similarity search in metric spaces," in *VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece*, M. Jarke, M. Carey, K. Dittrich, F. Lochovsky, P. Loucopoulos, and M. Jeusfeld, Eds.: Morgan Kaufmann, pp. 426-435, 1997.

[12] V. Dohnal , C. Gennaro , P. Savino , and P. Zezula "D-Index: Distance searching index for metric data sets," *Multimedia Tools and Applications*, vol. 21, pp. 9-33, 2003.

[13] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and Information Systems*, vol. 7, pp. 358-386, 2005.

[14] S. Belongie and J. Malik, "Matching with shape contexts," in *Proceedings : IEEE Workshop on Content-based Access of Image and Video Libraries, 2000.*, pp. 20-26, 2000.

[15] V. I. Levenshtein, "Binary codes capable of correcting deletions. Insertions and reversals," in *Soviet Physics Doklady*, vol. 10, no.8, pp. 707-710, 1966.

[16] L. Chen and R. Ng, "On the marriage of Lp-norms and edit distance," in *VLDB '04: Proceedings of the Thirtieth international conference on Very large data bases*. Toronto, Canada: VLDB Endowment, pp. 792-803, 2004.

[17] L. Devroye, L. Gyo"rfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*: Springer; Corrected edition (February 20, 1997), 1996.

[18] U. V. Luxburg and O. Bousquet, "Distance--based classification with Lipschitz functions," *Journal of Machine Learning Research*, vol. 5, pp. 669-695, 2004.

[19] R. Espinola and M. Khamsi, "Introduction to hyperconvex spaces," *Handbook of Metric Fixed Point Theory*, pp. 391-435, 2001.

[20] A. Guttman, "R-trees: a dynamic index structure for spatial searching," *SIGMOD Rec.*, vol. 14, pp. 47-57, 1984.

[21] E. Keogh and A. Ratanamahatana, "Everything you know about dynamic time warping is wrong," *3rd Workshop on Mining Temporal and Sequential Data, in conjunction with 10th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD-2004), Seattle, WA*, 2004.

[22] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, pp. 43-49, 1978.

[23] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 23, pp. 67-72, 1975.

[24] L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *The Journal of the Acoustical Society of America*, vol. 63, pp. S79, 1978.

[25] M. Watermann, T. Smith, and W. Beyer, "Some biological sequence metrics," *Advances in Mathematics*, vol. 20, 1976.

[26] E. Keogh, X. Xi, L. Wei, and C. A. Ratanamahatana, "The UCR time series classification/clustering homepage."