Correlation Based Dynamic Time Warping

Zoltán Bankó, János Abonyi

Department of Process Engineering, University of Pannonia, Veszprém, Hungary
www.fmt.uni-pannon.hu/softcomp, e-mail: abonyij@fmt.uni-pannon.hu

Abstract: Dynamic Time Warping (DTW) is a widely used technique for univariate time series comparison. This paper proposes a new algorithm for the comparison of multivariate time series which generalize DTW for the needs of correlated multivariate time series.

Keywords: multivariate time series, principal component analysis, dynamic time warping, segmentation

1   Introduction

A time series is a sequence of values measured as a function of time. These kinds of data are widely used in the field of process engineering, medicine, bioinformatics, chemistry and finance. The increasing popularity of knowledge discovery and data mining tasks for discrete data such as clustering (partitioning the data into coherent, but not predefined subsets), classification (placing the data into predefined, labeled groups) and kNN (k-nearest neighbor) search has indicated the growing need for similarly efficient methods for time series databases. These methods share a common requirement: a similarity measure has to be defined between the elements of a given database.

The similarity of multivariate time series can be approached from two different perspectives. The first way is the application of metrics based warping measures such as Dynamic Time Warping (DTW) and Longest Common SubSequence (LCSS). These techniques are perfectly suitable for univariate tasks like speech recognition, where the analyzed process represented by one variable only. In most cases these methods can be easily generalized for the needs of the multivariate time series, where the process depends on two or more variables. However, their application for comparing correlated multivariate time series is often not as effective as it is expected.

The direct comparison of the variables used by these approaches ignores the hidden process, i.e. the correlation between process variables and often this hidden process carries the real information. Hence, Principal Component Analysis (PCA) based similarity measures are used to overcome this problem. PCA considers the

time series as a whole but does not take into account the alterations in the relationship between the variables.

Our main goal is to construct a similarity measure which not only compares the variables directly, but the process behind them and is able to deal with the changes in the correlation structure of the variables.

1.1 Contribution

In this paper the problem of creating similarity measures for multivariate time series is investigated. We present a new and intuitive multivariate dynamic time warping method which is based on the correlation of variables and uses principal component analysis. This makes our similarity measure invariant to phase shifts of the time axis and it is capable to compensate the differences in sampling rates and 'locally elastic' shifts (local time warping) of the time series.

We do not examine the data validity and we omit the preprocessing steps for brevity. We treat that the time series are sampled (or can be resampled) at equispaced time points.

The rest of the paper is organized as follows. Section 2 details the theoretical background of the proposed similarity measure, i.e. dynamic time warping, principal component analysis and segmentation. In Section 3 we introduce our novel similarity measure which allows to combine the DTW and PCA, while Section 4 conducts a detailed empirical comparison of our method with basic techniques on verification databases widely used by the time series data mining community [1]. Finally in Section 6 we present our conclusions and suggestions for future work.

2 Background

An *n*-variable, *m*-element time series, $X_n = [x_1, x_2, \ldots, x_n]$, is an *m*-by-*n* element matrix, where $x_i = [x_i(1), x_i(2), \ldots, x_i(m)]^T$, is the *i*th variable and $x_i(j)$ its *j*th element. $X_n(j) = [x_1(j), x_2(j), \ldots, x_n(j)]$ is the *j*th sample of X_n.

The similarity between X_n and Y_n is denoted by $s(X_n, Y_n)$, where $0 \le s(X_n, Y_n) \le 1$, $s(X_n, Y_n) = s(Y_n, X_n)$ and $s(X_n, X_n) = 1$. Obviously, the similarity is nothing more than a real number between zero and one which expresses the tightness of connection between the processes behind the time series. The closer is the number to one the more similar the processes are. In

practice, the term distance or dissimilarity (d) is used instead of similarity. The value of the distance is given by a number ranged from zero to one. This can be associated with similarity: $d(X_n, Y_n) = 1 - s(X_n, Y_n)$. Note that is not expected from an applied distance to satisfy the triangle inequality when the distance is not used as the basic distance of an indexing method.

2.1   Dynamic Time Warping

Getting valuable information from time series data by data mining methods require a good similarity measure; even so, the traditional approaches are rarely precise enough for the most applications. This is caused by the brittleness of the conventional comparison methods. Distance measures such as Euclidean distance, are unable to handle the distortions in time axis, so these distortions almost randomly affect the distance between time series. The solution for this problem is the application of dynamic time warping which can 'warp' the original time series (nonlinearly dilate or compress their time axes) to be similar in shape to the query series as much as possible.

In the following the DTW algorithm is reviewed for univariate time series x and y. To align these sequences, firstly we define a grid D with size n x m where each cell represents a distance between the appropriate indices of the two time series. In this step we can choose any application-dependent distance like L_1 and L_∞ norms, but generally Euclidean distance is suggested because it allows of the efficient indexing of dynamic time warping [2]. Considering this we have:

$$D(i, j) = \sqrt[2]{(x(i) - y(j))^2} \tag{1}$$

Using grid D, we can create arbitrary mappings – called warping paths – between **x** and **y**. The construction of a warping path $[p(1), p(2), \ldots, p(l)]$ has to be restricted with the following constraints:

- *Boundary condition:* The path has to start in $D(1,1)$ and end in $D(n, m)$.

- *Monotonicity:* The path has to be monotonous, i.e. always heading from $D(1;1)$ to $D(n, m)$. If $p(k) = D(i, j)$ and $p(k+1) = D(i', j')$ then $i' - i \geq 0$ and $j' - j \geq 0$.

- *Continuity:* The path has to be continuous. If $p(k) = D(i, j)$ and $p(k+1) = D(i', j')$ then $i' - i \leq 1$ and $j' - j \leq 1$.
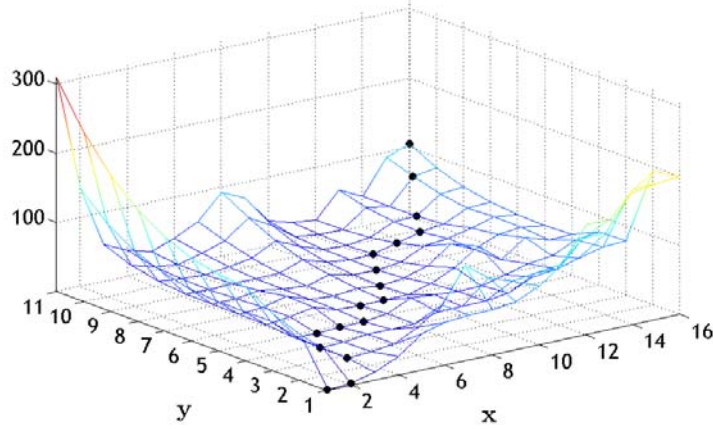
Figure 1
The cumulative distance matrix and the optimal warping path on it

To find the optimal warping path (the DTW distance of the two time series), every warping path has an assigned cost which is the sum of values of the affected cells divided by the normalization constant *K*:

$$d_{DTW}(x, y) = \min\left(\frac{\sum_{i=1}^{l} p(i)}{K}\right) \tag{2}$$

The value of *K* depends on the application and in most cases this is the length of the path, but it can also be omitted. More information about the method of defining *K* and its significance can be found in Reference [3]. Note that the Euclidean distance is a special case of DTW, i.e. we choose the path that is located on the diagonal of grid D and $K = 1$.

Obviously, the number of paths exponentially grows by the size of time series. Fortunately, the optimal path can be found in $O(nm)$ time with the help of dynamic programming using cumulative distance matrix $\hat{D}$. A cell of the cumulative matrix contains the sum of the appropriate cell value in matrix D and the minimum of the three cells from where the cell can be reached:

$$\hat{D}(i, j) = \min\begin{bmatrix} \hat{D}(i-1, j) + D(i, j) \\ \hat{D}(i, j-1) + D(i, j) \\ \hat{D}(i-1, j-1) + D(i, j) \end{bmatrix} \tag{3}$$

The DTW distance between the two time series can be found in $\hat{D}(n, m)$.

2.2   PCA Similarity Factor

PCA is a well known and widely used multivariate statistical technique for reducing the dimensionality of the data. Fundamentally, this is a linear transformation that projects the original (Gaussian distributed) data to a new coordinate system with minimal loss of information. In multivariate cases the information is the structure of the original data, i.e. the correlation between the variables and the alteration of the correlation structure among them. To create a projection based on this, PCA selects the coordinate axes of the new coordinate system one by one according to the greatest variance of any projection.

In projected space the first coordinate (first principal component) is the vector which has maximal variance in the original data space, the second coordinate (second principal component) has maximal variance in the original data space among all vectors which orthogonal to the first coordinate, the third coordinate orthogonal to the first two and has maximal variance in original data space, etc. Technically speaking, we select the eigenvectors of the covariance matrix of the original data (according to the descending order of corresponding eigenvalues) for axes in the new dimensional space.
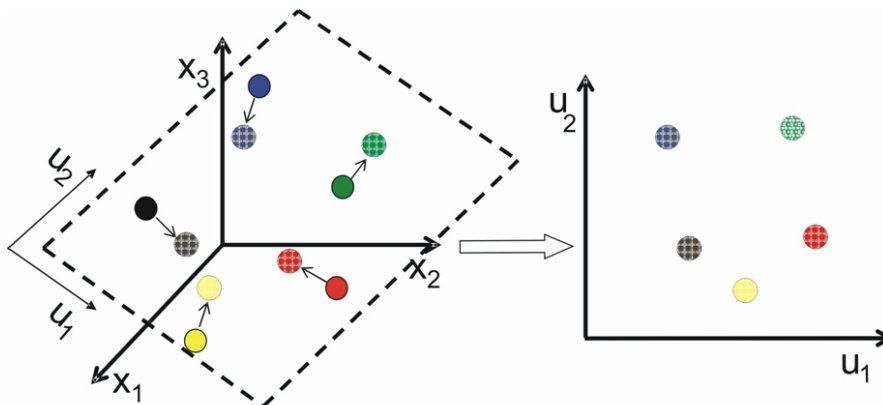

Figure 2

Dimension reduction with the help of PCA from 3 (filled dots) to 2 (patched dots) dimensions. Notice that the correlation between the dots is maximally preserved.

In most real multivariate processes the high-order principal components do not add any significant information to the low-order components so we can omit them to perform dimensionality reduction. Krzanowski defined [4] the PCA similarity factor to measure the similarity between different data by comparing the hyperplanes (the dimensionality reduced, new coordinate systems).

Let X_n and Y_n two multivariate time series with n variable. $U_{X_n,p}$ and $U_{Y_n,p}$ indicates the matrices of eigenvectors which belong the most important p eigenvalues of X_n's and Y_n's covariance matrix, i.e. the two hyperplanes. The similarity between them:

$$s_{PCA}(X_n, Y_n) = \frac{tr(U_{X_n,p}{}^T U_{Y_n,p} U_{Y_n,p}{}^T U_{X_n,p})}{p} \tag{4}$$

The similarity factor has a geometrical explanation, because it measures the similarity between the two hyperplanes by computing the squared cosine values between all the combinations of the first p principal components from $U_{X_n,p}$ and $U_{Y_n,p}$:

$$s_{PCA}(X_n, Y_n) = \frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{p} \cos^2 \Theta_{i,j} \tag{5}$$

where $\Theta_{i,j}$ is the angle between the *i*th principal component of X_n and the *j*th principal component of Y_n.

2.3   Segmentation

The *i*th segment of X_n is a set of consecutive time points, $S_i(a,b) = [X_n(a); X_n(a+1); \ldots; X_n(b)]$. The *c*-segmentation of time series X_n is a partition of X_n to c non-overlapping segments, $S_{X_n}^c = [S_1(1,a); S_2(a+1,b); \ldots; S_c(k+1,m)]$. In other words, a *c*-segmentation splits X_n to c disjoint time intervals, where $1 \leq a$ and $k \leq m$.

The main goal of segmentation is to find homogenous segments by the definition of a *cost* function. This function can be any arbitrary function which projects from the space of multivariate time series to the space of the positive real numbers and can be framed in several ways [5]. Abonyi et al. presented a PCA based segmentation in Reference [6], where the authors used the Hotelling's T^2 statistics and the Q reconstruction error as the measure of the homogeneity of the segments, i.e. the *cost* function.
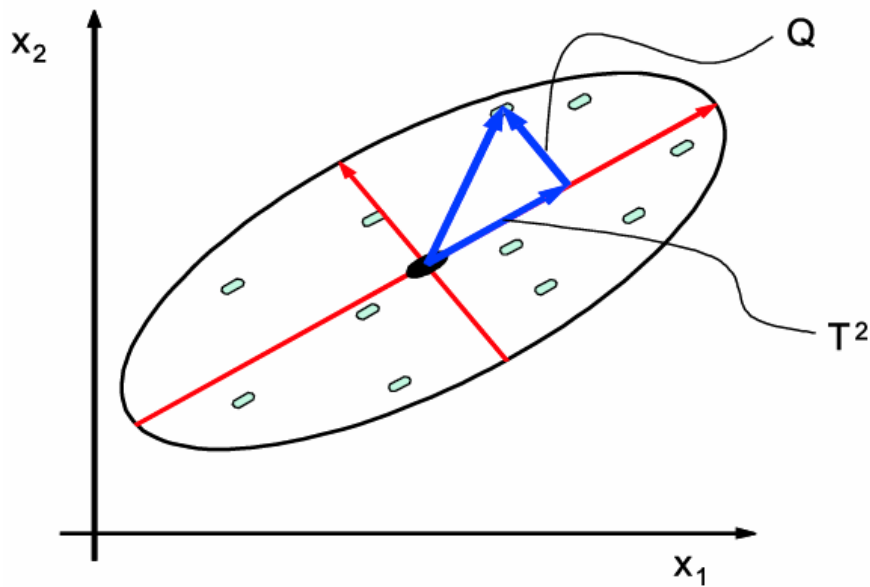
Figure 3
The distance measures used in PCA based segmentation

2.4 Related Work

The data mining community has been investigating different similarity measures for univariate time series for a long time. The majority of these works have focused on the Minkowski metrics (L_p norms), especially on the Euclidean distance. This popularity does not only arise from its simple geometric representation and fast computability, but also from the fact that the Euclidean distance is preserved under orthogonal transformations such as Fourier and Wavelet Transform and satisfies the triangle inequality.

From the viewpoint of data mining the main advantage of these transforms is that the original time series – in most cases – are well approximated using the first few coefficients of their transformed forms and the relations of the distances between the elements are also preserved. Technically speaking, the Euclidean distance allows the efficient indexing of large time series databases. The GEMINI framework generalized the creation of the indexing methods, thus, it provided facilities for new approaches which held out higher accuracy. In parallel with this, the increasing and less and less costly computational power made it possible to create similarity measures without indexing capabilities for tasks where the quality of the result is much more important than the speed of the calculation. Hence, new

techniques such as Dynamic Time Warping, Longest Common Subsequences and Symbolic Aggregate approXimation has been studied extensively from the requirements of ECG signal analysis through handwriting recognition and verification to tornado prediction.

On the other hand, the PCA based similarity factor also gains in popularity thanks to its outstanding property which makes possible to recognize the direction of the changes in the distance between the variables. Another advantage of this method is its optimal variable reduction behavior which makes it ideal for tasks with numerous variables such as process engineering jobs.

Johannesmayer [6] modified the PCA similarity factor by weighting each principal component with the square root of its eigenvalue:

$$s_{PCA}^{\lambda}(X_n, Y_n) = \frac{\sum_{i=1}^{p} \sum_{j=1}^{p} (\lambda_i^{X_n} \lambda_j^{Y_n}) \cos^2 \Theta_{i,j}}{\sum_{i=1}^{p} (\lambda_i^{X_n} \lambda_i^{Y_n})}, \tag{6}$$

where $\lambda_i^{X_n}$ and $\lambda_i^{Y_n}$ are the corresponding eigenvalues for the ith and jth principal component of X_n and Y_n.

This principle was developed by K. Yang [7] who presented the logical extension of PCA similarity factor and s_{PCA}^{λ} called Eros (Extended Frobenius Norm):

$$s_{Eros}(X_n, Y_n) = \sum_{i=1}^{n} w_i \left| U_{X_n}(i) U_{Y_n}^{T}(i) \right| = \sum_{i=1}^{n} w_i \left| \cos \Theta_i \right|, \tag{7}$$

where $\cos \Theta_i$ is the angle between the two corresponding principal components and w_i is the weighting vector which is based on the eigenvalues of the sequences in the data set.

3   Correlation Based Dynamic Time Warping of Multivariate Time Series

In the previous chapters, all of the required techniques have been reviewed for correlation based dynamic time warping of multivariate time series.

3.1 Proposed Algorithm

The following steps have to be performed for the realization of correlation based multivariate time warping:

- Segment the time series of the given database by correlation to create the most homogenous segments. The projection error or Hotelling's statistic can be used as the cost function. This segmentation can be done off-line in most cases.

- Segment the query time series when it arrives similarly.

- Calculate the distance between the query and the time series of the database with dynamic time warping. The base distance of DTW can be any distance which is based on the previously presented covariance based similarities (e.g. $1 - s_{PCA}$). The size of matrix $\hat{D}$ (the computation time of DTW) depends on the number of the segments.

4 Evolution Method and Experimental Results

A modified leave-one-out k-nearest neighbor search algorithm [8] was used to perform the validation of multivariate dynamic time warping. After the algorithm executes the recall-precision graph, which has been frequently used to measure the performance of Content Based Image Retrieval (CBIR) systems as well as Information Retrieval (IR) systems, can be plotted. The precision expresses the proportion of the relevant sequences from all the sequences are retrieved. Similarly, the recall is the percentage of the total relevant documents in a database retrieved by the k-nearest neighbor search.

4.1 Comparison of Similarity Measures

According to Reference [9], two dataset were used for the validation and both of them are available from the Internet. The first one is the database of SVC2004 (Signature Verification Conference 2004) which contains 40 sets of signature data[1]. This is an ideal dataset for correlation based multivariate time warping because the correlation alternates many times between the variables.

---

[1] Each set contains 20 genuine signatures from one signature contributor and 20 skilled forgeries from five other contributors. Although both genuine signatures and skilled forgeries are available, obviously the 800 genuine signatures were used for the validation.
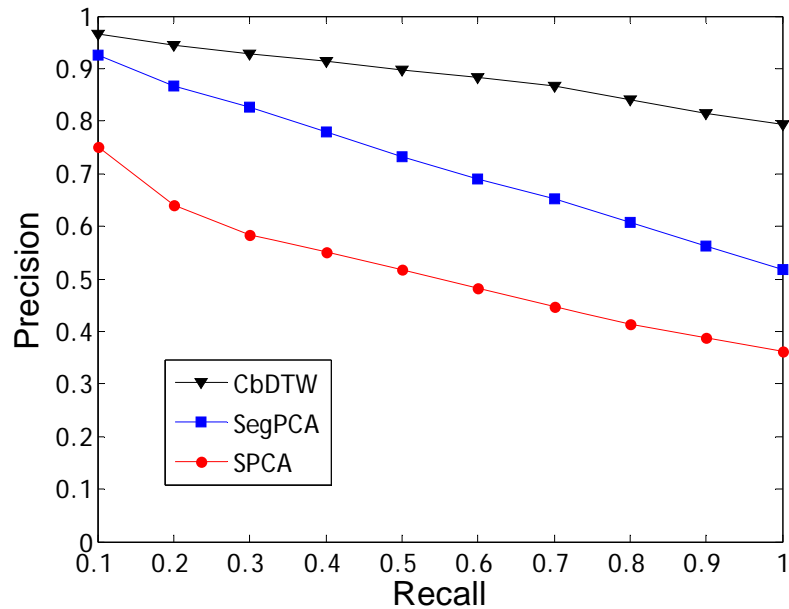
Figure 4
The Recall-Precision graph of SVC2004 dataset

Firstly, the best projection was selected for similarity factor based comparison. The best result was the one which compares the 2 dimensional hyperplanes. The first two eigenvalues describe more than the 99% of the total variance. This case is signed by the red line. After this, the sequences were segmented into 20 parts and the distances between the corresponding segments were calculated with the help of the Krzanowski similarity factor and the segmentation cost was the Hotelling's T^2 statistics. The graph of SegPCA shows that this technique really outperforms the segmentation free version. For fairness to this method, the correlation based DTW (CbDTW) used the same segmentation while performed the non-linear pairing between the segments. The superiority of this method is obvious for this dataset.
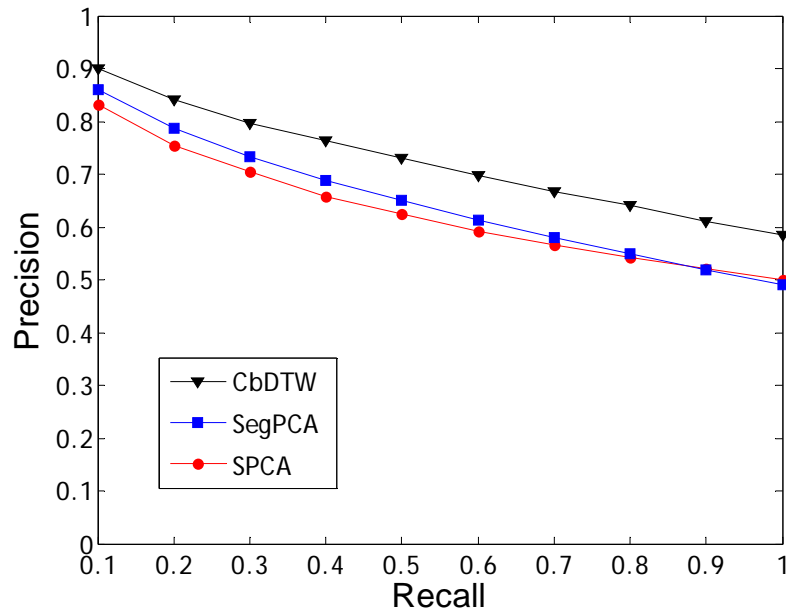
Figure 5

The Recall-Precision graph of AUSLAN dataset

The second validation was processed on the Australian Language Sign dataset (AUSLAN) collected by Mohammed Waleed Kadous [10]. This contains 95 signs and each of them has 27 examples, so the total number of sings is 2565. The average length of the time series in the database is only 57.

In Figure [5], the high values show that the correlation between the variables is much stronger than in SVC2004 dataset and this characterize the classes much more. The data was projected into 2 dimensional hyperplanes similarly to the first validation. These principal components describe more than the 95% of the total variance. For the last two methods, interpolation has to be executed and segmentation was applied on these interpolated versions of the original data. The segmentation cost was selected to Hotelling's T^2 statistic and the number of segments was 20.

Conclusions

The paper presented a new similarity measure for multivariate time series by combining dynamic time warping and principal component analysis. The results of the validations are more then promising, however, there is still a lot of work that has to be done and there are lots of questions that have to be answered. The other similarity measures (such as s_{Eros} and s_{PCA}^{λ}) do not provide as good results

as the PCA similarity factor. Presumably the reason for this is the segmentation. It has to be examined how the best segmentation result can be achieved for a given similarity measure.

Acknowledgement

References

[1] S. Hettich, S. D. Bay: The UCI KDD Archive, http://kdd.ics.uci.edu, University of California, Department of Information and Computer Science, 1999

[2] E. Keogh: Exact Indexing of Dynamic Time Warping, in Proceedings of 28th VLDB Conference, Hong Kong, China, 2002, pp. 406-417

[3] H. Sakoe, S. Chiba: Dynamic Programming Algorithm Optimization for Spoken Word Recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 27(1), 1978, pp. 43-49

[4] W. J. Krzanowski: Between-Groups Comparison of Principal Components, in Journal of the American Statistical Society, pp. 74:703-707, 1979

[5] E. Keogh, S. Chu, D. Hart, M. J. Pazzani: An Online Algorithm for Segmenting Time Series, in Proceedings of *ICDM*, 2001, pp. 289-296

[6] J. Abonyi, B. Feil, S. Németh, P. Árva: Principal Component Analysis-based Time Series Segmentation, in Proceedings of the IEEE International Conference on Computational Cybernetics, 2003

[7] M. C. Johannesmeyer: Abnormal Situation Analysis Using Pattern Recognition Techniques and Historical Data, M.sc. thesis, University of California, Santa Barbara, CA, 1999

[8] K. Yang, C. Shahabi: A PCA-based Similarity Measure for Multivariate Time Series, in Proceedings of the 2nd ACM International Workshop on Multimedia Databases, 2004, pp. 65-74

[9] E. Keogh, S. Kasetty: On the Need for Time Series Data Mining Benchmarks: a Survey and Empirical Demonstration, in Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 102-111

[10] Mohammed Waleed Kadous: Temporal Classification: Extending the Classification Paradigm to Multivariate Time Series, PhD thesis, School of Computer Science & Engineering, University of New South Wales, 2002