

Modeling Time Series Data for Supervised Learning

by

Mustafa Gokce Baydogan

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved August 2012 by the
Graduate Supervisory Committee:

George C. Runger, Chair
Robert Atkinson
Esma Gel
Rong Pan

ARIZONA STATE UNIVERSITY

December 2012

UMI Number: 3542183

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3542183

Published by ProQuest LLC (2012). Copyright in the Dissertation held by the Author.

Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against unauthorized copying under Title 17, United States Code



ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

ABSTRACT

Temporal data are increasingly prevalent and important in analytics. Time series (TS) data are chronological sequences of observations and an important class of temporal data. Fields such as medicine, finance, learning science and multimedia naturally generate TS data. Each series provide a high-dimensional data vector that challenges the learning of the relevant patterns

This dissertation proposes TS representations and methods for supervised TS analysis. The approaches combine new representations that handle translations and dilations of patterns with bag-of-features strategies and tree-based ensemble learning. This provides flexibility in handling time-warped patterns in a computationally efficient way. The ensemble learners provide a classification framework that can handle high-dimensional feature spaces, multiple classes and interaction between features. The proposed representations are useful for classification and interpretation of the TS data of varying complexity.

The first contribution handles the problem of time warping with a feature-based approach. An interval selection and local feature extraction strategy is proposed to learn a bag-of-features representation. This is distinctly different from common similarity-based time warping. This allows for additional features (such as pattern location) to be easily integrated into the models. The learners have the capability to account for the temporal information through the recursive partitioning method.

The second contribution focuses on the comprehensibility of the models. A new representation is integrated with local feature importance measures from tree-based ensembles, to diagnose and interpret time intervals that are important to the model.

Multivariate time series (MTS) are especially challenging because the input consists of a collection of TS and both features within TS and interactions between TS can be important to models. Another contribution uses a different representation to produce computationally efficient strategies

that learn a symbolic representation for MTS. Relationships between the multiple TS, nominal and missing values are handled with tree-based learners.

Applications such as speech recognition, medical diagnosis and gesture recognition are used to illustrate the methods. Experimental results show that the TS representations and methods provide better results than competitive methods on a comprehensive collection of benchmark datasets. Moreover, the proposed approaches naturally provide solutions to similarity analysis, predictive pattern discovery and feature selection.

PREVIEW

PREVIEW

To my family

ACKNOWLEDGMENTS

I want to express my deep and sincere gratitude to my advisor, Dr. George Runger, for his encouragement and generous support throughout my study at ASU. He offered me great research opportunities, resources, and trust which allowed me to fully explore the research area. Without his brilliant guidance, this dissertation would not have been possible.

I would also like to convey thanks to my dissertation committee: Dr. Robert Atkinson, Dr. Esma Gel, and Dr. Rong Pan for their valuable comments on this dissertation. I also would like to extend my gratitude to my industry collaborators: Dr. Eugene Tuv, and Dr. Ben Nelson.

My family supported me in every step I have taken since the very beginning of my graduate studies. I am mostly grateful to my father Mehmet Baydogan, my mother Perihan Baydogan, my sister Banu Gokcen Baydogan for their love and support. Without them, this work could not have been completed.

I am deeply indebted to my friends, Baykal Hafizoglu, Nedim Yel, Muhsin Menekse, Ahmet Cemal Durgun, Kerem Demirtas, Aysegul Demirtas and Tulin Inkaya, for their help, stimulating suggestions and encouragement. My special thanks go to my friends Mustafa Yuksel and Caglar Ata for their support through the course of this dissertation. Thanks for giving me a shoulder to lean on whenever I need.

Finally, my recent happiest moments were all with you, Didem Yamak. Thank you for your love, trust, and understanding. Thank you for providing me the continued moral support and encouragement to pursue my dreams. I am grateful to our journey so far and I am excited about our adventure ahead.

TABLE OF CONTENTS

	Page
LIST OF TABLES	ix
LIST OF FIGURES	xi
CHAPTER	
1 INTRODUCTION	1
1. A Bag-of-Features framework to classify time series	6
2. Supervised time series pattern discovery through local importance	8
3. Multivariate time series classification with learned discretization	9
4. Contributions	10
5. Organization of this dissertation	11
2 BACKGROUND	13
1. Notation	13
2. Dynamic time warping	17
3. Bag-of-Features approach	19
4. Random forest	21
3 A BAG-of-FEATURES FRAMEWORK TO CLASSIFY TIME SERIES	24
1. Abstract	24
2. Introduction	25
3. Background	30
4. Time Series Classification with a Bag of Features	33
4.1. Subsequences and feature extraction	34
4.2. Codebook and learning	37
4.3. Illustrative Examples	41

CHAPTER	Page
5. Experiments and Results	47
5.1. Classification accuracy	48
5.2. Computational complexity	56
6. Discussion	59
6.1. What OOB error rates provide	59
6.2. Shapelets and TSBF	60
7. Conclusions	63
4 SUPERVISED TIME SERIES PATTERN DISCOVERY THROUGH LOCAL IMPORTANCE	64
1. Abstract	64
2. Introduction	65
3. Background	71
3.1. Random Forest	71
3.2. Local importance measure from random forests	73
3.3. Tree models with interval features	75
3.4. Shapelets	78
4. Supervised Time Series Pattern Discovery through Local Importance . . .	79
4.1. Region of Interest Selection based on Local Importance	79
4.2. Pattern Discovery and Classification	84
4.3. Feature selection and summary using TS-PD	88
4.4. Parameters of TS-PD	90
5. Experiments	94
5.1. Computational accuracy	96

CHAPTER	Page
5.2. Computational complexity	100
5.3. Complexity reduction	106
6. Discussion	108
6.1. Illustrative example	108
6.2. Interpretability	110
6.3. Gesture recognition: an application of TS-PD to multivariate time series classification	112
6.4. Logical-Shapelets and TS-PD	114
7. Conclusion	115
5 MULTIVARIATE TIME SERIES CLASSIFICATION WITH LEARNED DIS- CRETIZATION	122
1. Abstract	122
2. Introduction	123
3. Background	131
4. Approach	132
4.1. Time Series Discretization using Tree-Based Classifiers	133
4.2. Classification	135
5. Experiments and Results	136
5.1. Univariate Time Series	137
5.2. Multivariate Time Series	145
6. Description of MTS datasets	147
6.1. Arabic speech recognition	148
6.2. Japanese Vowels	148

CHAPTER	Page
6.3. Pen-Based recognition of handwritten digits	149
6.4. ECG	149
6.5. Robot execution failures	149
6.6. Wafer	150
6.7. Australian sign language (AUSLAN)	150
6.8. Brazilian sign language (LIBRAS)	150
6.9. Character trajectories	151
6.10. Motion recognition-CMU_MOCAP_S16	151
6.11. Gesture recognition-uWaveGestureLibrary	151
6.12. Sensitivity Analysis	152
6.13. Computational Time Analysis	152
7. Conclusion	154
6 CONCLUSIONS AND FUTURE WORK	159
1. Conclusions	159
2. Future Work	161
2.1. Local feature extraction	161
2.2. Absence of the label information	162
2.3. Beyond time series	162
2.4. Similarity kernels	163
REFERENCES	164

LIST OF TABLES

Table	Page
1 Average test error rates over 10 replications for TSBF, TSBF without the subsequence location features (TSBF w/o location) and baseline RF classifier applied to two synthetic datasets.	44
2 Average test error rates over 10 replications for TSBF, and RF classifiers trained on an unsupervised codebook generated by K -means clustering applied to two synthetic datasets.	46
3 Characteristics of the time series datasets.	47
4 Parameter settings of TSBF	49
5 Error rates of TSBF for four different settings of z based on average, maximum and minimum of 10 replications, nearest-neighbor classifiers with dynamic time warping distance	52
6 Computation times of TSBF for different parameter settings.	57
7 Test and OOB error rates for different settings of maximum subsequence length.	60
8 Error rates of Logical-Shapelets and TSBF on 8 datasets.	62
9 Parameters of TS-PD	91
10 The OOB and test error rates of RF_{int} on CBF dataset for different interval settings.	93
11 Characteristics of the datasets	95
12 Error rates of TS-PD ($w = 6, 2000$ trees) for different settings of L	97
13 Error rates of TS-PD ($w = 6, 2000$ trees) for different settings of L (continued)	98

Table	Page
14	Computation times of TS-PD ($w = 6$, 2000 trees) for different settings of L 101
15	Computation times of TS-PD ($w = 6$, 2000 trees) for different settings of L (continued) 102
16	Error rates and computation times of TS-PD ($w = 6$, $L = 4$, 1000 trees) for different training data sizes. 107
17	Error rates of Logical-Shapelets and TS-PD 115
18	Sample database with 3 MTS from 2 classes 134
19	Parameter settings of TSBF 136
20	Characteristics of the univariate time series 139
21	Selected parameters based on OOB error rates. OOB error and test error rates of S-MTS 143
22	Selected parameters based on OOB error rates. OOB error and test error rates of S-MTS (continued) 144
23	Characteristics of MTS 145
24	Cross-validation error rates for S-MTS (10 replications) 147
25	Test error rates for S-MTS (10 replications) 148

LIST OF FIGURES

Figure	Page
1 Euclidean and Dynamic Time Warping distance computation [1]	4
2 Two time series from each class are shown ($T = 400, C = 2, y^n \in \{0, 1\}$) .	13
3 Two intervals (right) ($w = 50$) extracted from the time series (left)	14
4 A subsequence starting at time $t = 200$ (right) consists of $d = 5$ intervals of length $w = 20$ time units	15
5 Intervals of length $w = 40$ segmented from the time series using a sliding step of $r = 20$	16
6 A pattern of time series x^n composed of 3 discontinuous intervals.	16
7 Time alignment of two time-dependent sequences [2]	17
8 Cost matrix of two time series using the Manhattan distance [2]	18
9 Optimal warping path p^* , cost matrix c and accumulated cost matrix D . . .	23
10 Four steps to compute the bag-of-words representation for images [3] . . .	23
11 Instances from the OSUleaf dataset.	31
12 Generic description of the time series classification with a bag-of-features (TSBF) algorithm.	34
13 Interval and subsequence generation and representation.	37
14 More specific description of the time series classification with a bag-of- features (TSBF) algorithm.	40
15 The number of peaks example time series	42
16 Distributions of the subsequences in the feature spaces of interval means for two examples	50
17 One time series from each class with peak location	51

18	The average OOB error rates on the training data of RFts and RFsub over all datasets.	51
19	Scatter plot of error rates from TSBF and NNDTWNoWin.	53
20	Scatter plot of error rates from TSBF and NNDTWBestWin.	54
21	Boxplot of the replication results for TSBF ($z = 0.5$).	55
22	Computation time of TSBF over all datasets for all z settings	58
23	OOB error rates of TSBF ($z = 0.5$) with $b = 10$ and with $b = 50$	61
24	Two sample time series from different classes.	67
25	Training time series instances from CBF dataset.	76
26	Decision trees built using C4.5 [4] on the interval features.	77
27	Illustration of the classes for Gun-Point dataset.	79
28	Illustration of feature generation on the intervals of one time series from CBF dataset.	81
29	Three time series from CBF dataset and corresponding local importance plot.	83
30	Normalized local importance information on CBF dataset and time series of each class	84
31	Illustration of distance computation over the time series for a generated patterns	86
32	Variable importance of <i>RFpattern</i> based on Gini measure on CBF dataset	89
33	First 12 important patterns of TS-PD for CBF dataset	90
34	The OOB error rates of <i>RFint</i> and <i>RFpattern</i> of CBF dataset	92
35	Progress of OOB error rates and test error rates over L settings.	94
36	Scatter plot of error rates of TS-PD vs NNDTWNoWin and NNDTWBestWin.	99

37	Training and testing times of TS-PD on Two Patterns dataset for increasing dataset sizes	104
38	Training and testing times of TS-PD for FacesUCR dataset for different w and L settings.	106
39	Training and testing times for series of different length	106
40	Illustration of the transformation of a face image to the time series.	108
41	The progress of the OOB error rate of RF_{int}	109
42	Normalized local importance information on FacesUCR and time series of each class	116
43	The OOB error rates of $RF_{pattern}$ over trees for $w = 20, L = 4$	117
44	First five important patterns of TS-PD (Gun-Point dataset)	118
45	First five important patterns of TS-PD (Sony AIBO Robot)	119
46	First five important patterns of TS-PD (Coffee)	120
47	Univariate representation of the accelerometer data.	120
48	Gesture vocabulary from [5] and important patterns	121
49	SAX representation with a word size of 8 and alphabet size of 3	126
50	Alternative representations for MTS	127
51	One time series of each class from CBF dataset.	140
52	The feature space and the partitions (symbols) from the decision tree	141
53	A visual example of the representation based on symbol frequencies	142
54	Boxplot of OOB error rates and test error rates for each combination setting over multiple trees for Non-Invasive Fetal ECG Thorax1 dataset	156
55	The mean computation times with changing R and J_{ins}	157

56	The mean computation times with changing the number of training instances and time series lengths	157
57	The boxplot of the computation times of S-MTS with changing the number of variables	158

PREVIEW

CHAPTER 1

INTRODUCTION

In the last decade, the increasing use of temporal data, especially time series data, has initiated a great deal of research and development attempts in the field of data mining. Time series data which is chronological sequences of observations is one of the important class of temporal data. Many data sources in different fields, such as in medicine, finance, multimedia and learning sciences naturally generate time series data. For example, an ElectroCardioGram (ECG) is used to identify temporal patterns in heart signals to identify abnormal heart rhythms [6]. Average electrical voltage produced by the beating of the hard muscle is measured over the human body. An ECG is visualized as a 2D plot, where x axis is the time and y axis is the average voltage measured by the electrodes. In the field of seismology, seismograms are used to identify seismic events. A seismogram is a record of the ground motion produced by an earthquake, explosion, or other ground-motion sources [7]. The ground motion is identified by a seismograph at a measuring station as a function of time. Nowadays, Electroencephalography (EEG) which is the recording of electrical activity along the scalp is used to understand the brain activity and connectivity under different experimental conditions. EEG visualizes the voltage fluctuations resulting from ionic current flows within the neurons of the brain over the time.

Time series data is characterized by its numerical and continuous nature [8]. Time series are considered as a whole instead of individual numerical fields because of the temporal ordering in the data. This makes time series analysis different from other data analysis problems, in which there is no natural ordering of the observations. Moreover, another problem is that each series provide a high-dimensional data vector that challenges the analysis. The high-dimensionality can be handled by dimensionality reduction techniques such as feature selection when the temporal ordering is not important. However, entire series should

be considered as a vector in time series analysis problems since the relations between the certain time points may be of interest. Therefore, traditional dimensionality reduction techniques may not work well for the time series data. Real-world time series data is often high-dimensional, contains nonlinear relationships between its variates, and has long-range dependencies. Due to these complexities, time series data mining has received great interest over the past decade.

Time series data mining approaches focus on various problems. The major tasks considered in this context are pattern discovery and clustering, classification, rule discovery and summarization [8]. Although these tasks are presented separately, they are not independent. For instance, clustering result on time series may be useful to a classification task. Therefore, a study on one particular task may provide solutions to other tasks.

A fundamental problem in time series data mining approaches is how to represent the time series data. The representation is important to discover the useful information from the high-dimensional data efficiently rather than analyzing or finding statistical properties on the whole series. High-level representation of the original raw data is generally used as a feature extraction step, or simply to make the storage, transmission, and computation of massive dataset feasible in these approaches [9]. The time series representation strategies are categorized into two classes [9]: data adaptive (adaptive basis representation) and nondata adaptive (fixed basis representation). Examples of data adaptive approaches are Singular Value Decomposition (SVD) [10], Piecewise Linear and Piecewise Constant models (PAA) [11] and Symbolic Aggregate Approximation (SAX) [12]. Nondata adaptive approaches represent the time series in the transformation domain using mostly Discrete

Fourier Transform (DFT) [13] and Discrete Wavelet Transform (DWT) [14]. This thesis explores new adaptive basis representations for time series classification.

Time series classification is a supervised learning problem in which the input consists of a set of training examples and associated class labels, where each example is formed by one or more time series (variables) and the aim is to label test examples to predefined classes. Time series classification is an important task with many challenging applications including finance, science, natural language processing and medicine. For example, a cardiologist might be interested in analysis of ECG signals from different patients in order to see whether a particular patients, e.g., patients with a history of some disease, have different temporal s in their heart signals than a control group [6]. Seismologist aim at discriminating the nature of the seismic waves to classify events such as earthquakes, mining explosions or nuclear explosions [7]. Moreover, EEG records are used in a learning environment to understand the perceived difficulty by classifying the EEG signals based on the puzzle difficulty. Effective and efficient data mining methods are required for the knowledge extraction in such applications.

The algorithms proposed for time series classification can be divided into instance-based and feature-based methods in general. Instance-based classifiers predict a test instance based on its similarity to the training instances. For example, nearest neighbor (NN) classifiers classify objects based on the closest training examples in the feature space and one-nearest-neighbor classifiers with Euclidean (NNEuclidean) or a dynamic time warping distance (NNDTW) have been widely, and successfully used [15–19] in time series classification.

One-nearest-neighbor (NN) classifiers with Euclidean distance do not work well if the patterns of interest translate or dilate over time. DTW [20] is a method that allows a measure of the similarity of time series independent of certain non-linear variations in the time dimension. The idea of DTW is illustrated in Figure 1. Euclidean distance is computed by matching the observation at the same time points. Conversely, DTW aligns the observations using a dynamic programming approach that maximizes the similarity of the time series while satisfying the time ordering of the observations. Therefore, DTW recognizes the similarity of the time series better than the Euclidean distance.

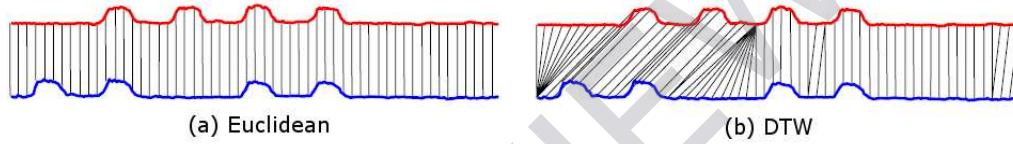


Figure 1. Euclidean and Dynamic Time Warping distance computation [1]. The grey lines indicate that distance is computed over the observations at either end of the line. Alignment of two time series by DTW recognizes the similarity of the series better than the Euclidean Distance

The majority of the NN classifiers works on the raw (observed) data. On the other hand, there are studies based on alternative time series representations. These studies search for similarity on features instead of the raw data. For example, Symbolic Aggregate Approximation (SAX) [12] basically represents the time series based on the mean level of the intervals extracted from the time series. An NN classifier based on this representation searches for similarity on the mean feature of the intervals. We consider the most accurate NN classifiers based on the raw data in this thesis.

NN classifiers with appropriate distance measures are known to provide strong and robust solutions [21, 22] although their space and time requirements may be problematic for

some application. NN classifiers are easy to understand and do not require the setting of many parameters, but they typically do not provide insight into time series features important to the classifier. Why a particular instance is assigned to a certain class is not clear.

Feature-based classifiers work on the features of the time series to reduce the dimensionality. They are interpretable and generally faster than instance-based classifiers depending on the feature extraction method and classification algorithm. The feature extraction step should handle the temporal information relevant to classification and a classifier that can take the temporal relations into account is required. Two types of features are generated in these approaches, global and local features. Global features are extracted from each time series and provide a compact representation of the time series (such as the mean of all observed values) but they are usually insufficient to represent time series information useful to classifiers. On the other hand, local features are extracted from segments of the time series and require such segments to be determined. Since the set of local features may vary in cardinality and lack a meaningful ordering, many classification algorithms requiring feature vectors of fixed dimension have problems in handling the local feature set.

In this thesis, we explore the problems related to time series classification. We propose time series representations that overcome some limitations of existing approaches for classifying the time series. In particular, we consider the following questions in details:

- Long time series with time warped patterns, relatively short features of interest, and moderate noise, are difficult to identify. What are the benefits of the feature-based approaches in such cases? Are there methods that can handle time warping with all the benefits of a feature-based approach?

- Why is a time series assigned to a certain class? Are there patterns specific to certain classes? Which patterns are relevant to the classification task?
- There might be more than one time series relevant to the classification task and multiple series challenge the similarity-based approaches. Scalability of the approaches become important as the number of time series increases. Also, both features within the time series and interactions between the time series can be important to models. Are there computationally efficient strategies to learn both relations simultaneously for time series classification?

1. A Bag-of-Features framework to classify time series

A framework based on the bag-of-features (BoF) representation is proposed to benefit from the speed and other advantages of feature-based methods to handle the problems for which NN classifiers with DTW distance are challenged. A BoF representation characterizes complex objects by feature vectors of sub-objects. We propose interval selection and local feature extraction strategies to explore time series representation that can handle translation and dilations based on the BoF idea.

To capture local information, random subsequences are extracted from each time series and further divided into intervals. The subsequences vary randomly in length and location. The number of intervals that partition a subsequence are fixed so that the interval length varies with the subsequence length. Several features (such as the mean, standard deviation, etc.) are extracted from each interval and these features comprise a row in a new data matrix X (one row for each subsequence). Because the subsequences selected vary in length and location, a particular column in X consists of features from different time locations computed over different length intervals. Consequently, the similarity between time series

can be captured independent of certain non-linear variations in the time dimension. This representation captures information in a manner similar to DTW, but from a very different construction. After representing the features of the subsequences in data matrix X , a classifier is trained assuming that each subsequence has the label of the time series from which it is extracted. Classification results on the subsequences are summarized to obtain the new representation for the time series. This data structure along with a tree-based ensemble allows for relevant features to be used by the classifier while irrelevant ones tend to be ignored.

Our local feature generation scheme allows for a novel representation that captures information in a manner similar to DTW, we then label the subsequences and use a supervised approach to summarize the local information unlike the existing studies. Our supervised approach allows for desirable properties for time series classification problem. It provides fast and efficient time series representation for classification even with very basic features such as slope, mean and variance from the subsequences. Global features (e.g autocorrelation of the time series) can also be extracted from the time series and combined with other features. Finally time series may be classified via any supervised learner. We denote the new algorithm as BoF framework to classify Time Series (TSBF).

In Chapter 3, we will address time series classification problem based on bag-of-features representation. We show how TSBF handles the temporal data and demonstrate its efficiency and accuracy by comparing to alternative time series classifiers on a full set of benchmark data sets.

2. Supervised time series pattern discovery through local importance

In Chapter 4, we consider a framework for finding important patterns of time series for classification. We focus on finding the segments of the time series that have potential to distinguish the classes. These segments are referred as the regions of interest. Regions of interests are very important to understand the temporal relations. Moreover, they help to reduce the effort in searching for the time segments useful to a classifier. After finding the region of interests for each time series, we generate sequences from these regions. These sequences are referred as patterns. We generate multiple patterns from the time series and find the best matching segments of the time series to these patterns. Then each time series is represented by the distances of the patterns to the best matching segments of the time series. Another classifier is then trained on this representation. A feature selection algorithm on the new feature set allows for finding the patterns that are critical in classification.

A feature-based algorithm is used to reduce the effort to prune the search space of the regions of interest in our algorithm. [23] also discusses the necessity of pruning the search space to find the regions relevant to classification and proposes a distance-based method. Feature-based approaches allow for some desirable properties such as handling the interactions and fast computation. Interaction between the features in this context is the relationship of the patterns over multiple intervals that may define a class as discussed by [23].

In Chapter 4, we will describe how the interpretability is achieved through the pattern discovery process. We illustrate the compactness of the new representation which reduces the time and space required for classification.