

# Contents

Resumen . . . . .	i
Abstract . . . . .	ii
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	2
1.2 Goals . . . . .	3
1.3 Methodology . . . . .	3
1.3.1 Component Description . . . . .	3
1.3.2 Evaluation . . . . .	4
1.4 Research Problems . . . . .	5
1.4.1 Time Series Characterization . . . . .	5
1.4.2 Efficiency in Anomaly Detection . . . . .	5
1.5 Publications . . . . .	6
1.6 Organization . . . . .	6
<b>2 Background</b>	<b>8</b>
2.1 Introduction . . . . .	8
2.2 Basic Concepts . . . . .	8
2.2.1 Time Series . . . . .	8
2.2.2 Primary Data Mining Tasks . . . . .	9
2.3 Distance Measures for Time Series . . . . .	12
2.3.1 Main Challenges in the Matching . . . . .	12
2.3.2 Shape-based Distances . . . . .	12
2.3.3 Edit-Based Distances . . . . .	14
2.3.4 Distances for Multivariate Time Series . . . . .	15
2.4 Time Series Representation . . . . .	15
2.4.1 Feature Extraction Techniques . . . . .	16
2.4.2 Representation Models . . . . .	19
2.4.3 Representation of Multivariate Time Series . . . . .	21
2.5 Indexes for Time Series . . . . .	22
2.5.1 MBR-based Index . . . . .	23
2.5.2 SAX-based Index . . . . .	26
2.5.3 Indexes for Multivariate Time Series . . . . .	27
2.6 Anomaly Detection in Time Series . . . . .	28
2.6.1 Anomaly Definitions . . . . .	28
2.6.2 Discord Discovery Approach . . . . .	29
2.6.3 Efficient Techniques . . . . .	30

2.6.4	Accuracy Measure for Anomaly Detection . . . . .	33
2.6.5	Efficient Model for Large Streaming Data . . . . .	33
2.7	Benchmarks for Time Series . . . . .	35
2.7.1	Datasets for Classification and Clustering . . . . .	35
2.7.2	Datasets for Anomaly Detection . . . . .	35
2.8	Summary . . . . .	36
<b>3</b>	<b>Similarity Search in Time Series using Feature Signature</b>	<b>39</b>
3.1	Introduction . . . . .	39
3.2	Signature Quadratic Form Distance (SQFD) . . . . .	40
3.3	Feature Signature for Time Series . . . . .	41
3.3.1	Local Features Extraction . . . . .	41
3.3.2	Clustering . . . . .	43
3.3.3	Local Aggregation . . . . .	43
3.3.4	Feature Signature Matching . . . . .	44
3.3.5	Additional Optimizations . . . . .	45
3.4	Experimental Evaluation . . . . .	45
3.4.1	Performance of Local Features . . . . .	46
3.4.2	Performance of Feature Signature . . . . .	48
3.5	Summary . . . . .	54
<b>4</b>	<b>Discord Discovery for Non-Normalized Time Series</b>	<b>55</b>
4.1	Introduction . . . . .	55
4.2	Techniques Based on Bounding Boxes . . . . .	57
4.2.1	Array of MBRs . . . . .	57
4.2.2	List of MBRs . . . . .	58
4.2.3	Discord Discovery Heuristics . . . . .	59
4.3	Online Detection . . . . .	61
4.4	Experimental Results . . . . .	64
4.4.1	Effectiveness of Time Series Transformations . . . . .	64
4.4.2	Structural Evaluation . . . . .	65
4.4.3	Efficiency Comparison . . . . .	67
4.4.4	Evaluation of the Online Algorithm . . . . .	69
4.5	Summary . . . . .	71
<b>5</b>	<b>A Multi-resolution Approximation for Time Series</b>	<b>72</b>
5.1	Introduction . . . . .	72
5.1.1	Background . . . . .	73
5.2	Multi-resolution Trend-Value Approximation . . . . .	74
5.2.1	Motivation . . . . .	74
5.2.2	Bottom-up Construction Algorithm . . . . .	76
5.2.3	Distances Measures . . . . .	77
5.2.4	Symbolic Representation . . . . .	78
5.2.5	Indexing . . . . .	79
5.3	Multi-resolution Discord Discovery . . . . .	81
5.3.1	Building Algorithm . . . . .	81
5.3.2	Discord Discovery Heuristics . . . . .	83

5.4	Experimental Results . . . . .	84
5.4.1	Classification . . . . .	84
5.4.2	Anomaly Detection . . . . .	88
5.5	Summary . . . . .	90
<b>6</b>	<b>Discord Discovery for Multivariate Time Series using Metric Indexes</b>	<b>92</b>
6.1	Introduction . . . . .	92
6.2	Metric Indexes . . . . .	94
6.2.1	Pivot-based Index . . . . .	94
6.2.2	List of Clusters . . . . .	97
6.2.3	Locality Sensitive Hashing (LSH) . . . . .	100
6.2.4	Snake Table . . . . .	103
6.3	Experimental Results . . . . .	104
6.3.1	Structural Evaluation . . . . .	105
6.4	Summary . . . . .	109
<b>7</b>	<b>Conclusions and Further Work</b>	<b>112</b>
7.1	Summary of Contributions . . . . .	112
7.2	Future Work . . . . .	113
	<b>Bibliography</b>	<b>115</b>

# List of Tables

- 2.1 Comparison of distance measures regarding their robustness. . . . . 15
- 2.2 Related works for each anomaly definition. . . . . 28
- 2.3 Benchmarks for time series and some applications topics. . . . . 35
- 2.4 UCR Time Series Archive: data collections used for our assessments. . . . . 37
- 2.5 Datasets for discord discovery evaluation. . . . . 38
  
- 3.1 Heuristics for generating the feature signature. . . . . 44
- 3.2 Computational complexity for the ED and the DTW. . . . . 45
- 3.3 Time series datasets for evaluating the feature signature. . . . . 46
- 3.4 Classification error before and after setting the referential position. . . . . 50
- 3.5 Performance of all proposed heuristics for generating the signature. . . . . 51
- 3.6 Comparison of our approach with other state-of-the-art feature models (using the NN Classifier). . . . . 53
  
- 4.1 Effectiveness of different subsequence transformations. . . . . 65
  
- 5.1 Notation used in this chapter. . . . . 75
- 5.2 Classification error using the best configuration of three numeric representations. . . . . 86
- 5.3 Equalizing the quantitative information. . . . . 86
- 5.4 Classification error using raw representation and MTVA representation. . . . . 88
- 5.5 Percentage of true detections for trend-value representations. . . . . 89
- 5.6 Parameters to be tuned. . . . . 90
  
- 6.1 Percentage of true detections by different time series transformation. . . . . 94
- 6.2 Percentage of true detections by different time series representations. . . . . 95
- 6.3 Structural policies and parameters for each metric index. . . . . 105
  
- 7.1 Overall summary of contributions. . . . . 112

# List of Figures

1.1	Schematization of components involved in our proposal. . . . .	4
2.1	Three challenges for matching time series. . . . .	13
2.2	Two distance measures for time series: One-to-one matching by the Euclidean Distance (left) and finding the best alignment by the Dynamic Time Warping (right). . . . .	14
2.3	Dimensionality reduction with PAA technique. . . . .	17
2.4	An example of SAX representation for a time series $P$ : $SAX(P) = \{00, 11, 10, 01\}$ , where $d = 4$ and $\alpha = 4$ . . . . .	18
2.5	Interest points detection with PIP technique. . . . .	18
2.6	Time series classification using Bag of Words. . . . .	21
2.7	A filter-and-refine architecture for efficient search of time series. . . . .	23
2.8	R*-Tree index for time series using PAA descriptor. . . . .	24
2.9	Distance between a time series query and any MBR. MINDIST corresponds to the red lines. . . . .	25
2.10	Two indexing structures for time series using SAX descriptor. . . . .	27
2.11	Three different cases in measuring the accuracy of detection. . . . .	33
2.12	Two efficient models for anomaly detection on large streaming data. . . . .	34
3.1	Similarity search model using Local Features and Feature Signatures. . . . .	40
3.2	Methods for generating the feature signature. . . . .	42
3.3	Evaluation of time series representation models based on feature extraction. . . . .	47
3.4	Building time of the local features in NonInvasiveFatalECG_TORAX1 dataset. We normalize the building time in terms of the search time. . . . .	48
3.5	Distribution analysis of a sample of local features with a boxplot for each vector dimension. . . . .	49
3.6	Performance of the two signature generation methods increasing the number of centroids. . . . .	49
3.7	Evaluating the feature signature as a reduced model of the local features. . . . .	52
3.8	Searching time of the feature signature in the NonInvasiveFatalECG_TORAX1 dataset using the cDTW distance. We normalize the time in terms of raw representation. . . . .	52
3.9	Building time by five feature-based Models in the NonInvasiveFatalECG_TORAX1 dataset. . . . .	54

4.1	<b>Finding the discordant subsequence using different transformation techniques.</b> Note that normalization requires an appropriate $\varepsilon$ to minimize the effect of noisy subsequences. We display the original time series (top graphs) and the best non-self match distance of each subsequence (bottom graphs). . . . .	56
4.2	Array of MBRs for discord discovery. . . . .	57
4.3	List of MBRs for discord discovery. . . . .	59
4.4	MINDIST returns the minimum distance between a subsequence query $C_q$ and any subsequence bounded in $R$ , which is illustrated by the red lines. . . . .	60
4.5	Performance of discord discovery in a non-stationary ECG time series. . . . .	66
4.6	The number of computed distances by four split algorithms. . . . .	67
4.7	The number of computed distances and the building time (CPU runtime in seconds) by five center selection algorithms. . . . .	67
4.8	Performance of discord discovery techniques over non-normalized subsequences. . . . .	68
4.9	Performance of discord discovery techniques over normalized subsequences. HOT SAX* is an optimized version of the original HOT SAX algorithm. . . . .	69
4.10	Online anomaly detection over two real time series using a sliding window of size $w = 128$ and $w = 108$ , respectively. The red points define the anomalous region detected by our online algorithm. . . . .	70
4.11	The number of computed distances by both the nearest non-self match search and the anomaly detection algorithm. . . . .	71
5.1	Three techniques for time series representation based on trend and value: (a) Piecewise Trend Approximation, (b) Piecewise Trend-Value Approximation and (c) Extended Trend-Value Approximation. . . . .	74
5.2	A comparison of the ability of two time series representations to cluster five members of the CBF dataset using the Euclidean distance. . . . .	75
5.3	Construction of the Multi-resolution Trend-Value Approximation. . . . .	76
5.4	Density of the slope varying the level of resolution in ECG time series. . . . .	79
5.5	Index model for the MTVA representation. . . . .	80
5.6	Lower bounding trend-value cost. The blue line represents a trend-value pair stored in our database and the green line is the query. . . . .	81
5.7	Multi-resolution Index Model for the MTVA representation. . . . .	82
5.8	Comparison of our multi-resolution representation with other state-of-the-art techniques. . . . .	86
5.9	Efficiency improvement by the MTVA Distance regarding the classic ED that uses raw representation. . . . .	87
5.10	Efficiency improvement of our multi-resolution method in Anomaly Detection. . . . .	90
5.11	Structural comparison of two indexing methods for MTVA representation. . . . .	90
6.1	Anomaly detection in a ECG time series using sliding window of size 150. . . . .	93
6.2	Metric index based on pivots (left) and the condition of discarding objects in the search (right), where $r$ is the distance of the best match so far. . . . .	96
6.3	List of Clusters: A dynamic and compact partitioning algorithm for metric spaces. . . . .	99
6.4	LSH Index with different number of seeds. . . . .	101
6.5	Structural evaluation for the pivot-based index. . . . .	107

6.6	Structural evaluation for the snake table. . . . .	107
6.7	Structural evaluation for the list of clusters. . . . .	107
6.8	Structural evaluation for the LSH index. . . . .	108
6.9	The number of computed distances by some metric indexes in offline anomaly detection. . . . .	109
6.10	Anomaly detection in two multivariate time series using a sliding window of size $w = 256$ and $w = 150$ , respectively. The red points define the anomalous region detected by our online algorithm. . . . .	110
6.11	The number of computed distances by both the nearest non-self match search and the anomaly detection algorithm for several metric indexes. . . . .	110