Faculty of Engineering and Information Technology
University of Technology Sydney

Multivariate Sequential Contrast Pattern Mining and Prediction Models for Critical Care Clinical Informatics

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Shameek Ghosh

December 2017

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not been previously submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used have been reported in the thesis.

Signature of Candidate

_____

Acknowledgments

Foremost, I would like to express my deepest gratitude to my supervisor Prof. Jinyan Li for his continuous support to my doctoral study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance has helped me in learning to carry out strong and effective research and in the preparation of this thesis. I could not have imagined having a better mentor and advisor for my doctoral research.

Also, I would like to thank my co-supervisor Prof. Hung Nyugen, Dr. Mengling Feng, Prof. Ramamohanarao Kotagiri, and Prof. Longbing Cao for their continuous support and scientific advice during my research. Without their professional guidance, this thesis would not have been possible.

Additionally, I would like to thank my colleagues, Dr. Qian Liu, Jing Ren, Renghua Song, Yi Zheng, Chaowang Lan, Hui Peng, and Yuangsheng Liu for their strong support and numerous stimulating discussions.

Besides, I offer my regards to all of my co-workers at the Advanced Analytics Institute and Faculty of Engineering and IT, and thank them for their support in the completion of this dissertation.

Shameek Ghosh
December 2017, UTS

Contents

Certificate . . . . . **i**
Acknowledgment . . . . . **iii**
List of Figures . . . . . **ix**
List of Tables . . . . . **xi**
List of Publications . . . . . **xiii**
Abstract . . . . . **xv**

Chapter 1 Introduction **1**
1.1 Background 1
1.2 Mining of Sequential Patterns 4
1.3 Mining Useful Patterns for Critical Care Decision-support . . 6
1.3.1 Problem Statement 7
1.3.2 Research Motivations 8
1.4 Limitations and Challenges 10
1.5 Research Issues 12
1.6 Research Contributions 13
1.7 Thesis Structure 14

Chapter 2 Literature Review **16**
2.1 Frequent Pattern Mining Framework 16
2.1.1 Sequential pattern mining 19
2.2 Emerging Patterns 26
2.2.1 Estimating the quality of emerging patterns 33
2.2.2 Mining paradigms 34

2.3 Temporal Patterns 40
2.3.1 Substring patterns 43
2.3.2 Sequential patterns 43
2.3.3 Time-interval patterns 44
2.4 Pattern mining in Critical Care Applications 45
2.4.1 Short-term predictive modelling 46
2.4.2 Long-term predictive modelling 47
2.4.3 State-of-the art in ICU informatics 49
2.4.4 Research issues 50

Chapter 3 Hypotension Risk Prediction via Sequential Contrast Patterns of ICU Blood Pressure 52

3.1 Introduction 53
3.1.1 Aims of the study 54
3.1.2 Research contributions 55
3.2 Problem Definition 55
3.2.1 Formulation of the AHE prediction problem 56
3.2.2 Related works for prediction of hypotension 57
3.3 Methodology 60
3.3.1 Data extraction 60
3.3.2 Data discretization 63
3.3.3 Mining gap-constrained sequential contrast patterns . . 64
3.4 Prediction Results 68
3.4.1 Prediction performance on the two data sets 70
3.4.2 Discussion 72
3.5 Examples and Clinical Significance of Sequential Contrast Patterns 76
3.5.1 Sequential pattern examples 77
3.5.2 Pattern visualization and clinical interpretation 78
3.6 Conclusion 79

Chapter 4 Using Sequential Patterns as Classification Features for Accurate Prediction of ICU Events . . . 81
4.1 Introduction . . . 82
4.2 Related Work . . . 84
4.3 Methodology . . . 86
4.3.1 Data discretisation . . . 87
4.3.2 Mining sequential contrast patterns . . . 87
4.3.3 Integrating sequential patterns for model construction . 88
4.4 Results and Discussions . . . 91
4.4.1 Dataset description . . . 91
4.4.2 Classification results . . . 93
4.5 Conclusion . . . 98

Chapter 5 Septic Shock Prediction for ICU Patients via Coupled HMM Walking on Sequential Contrast Patterns . . . 99
5.1 Introduction . . . 99
5.1.1 Contributions . . . 102
5.2 Related Work . . . 102
5.2.1 Previous studies in septic shock prediction . . . 103
5.2.2 Pattern-based classification models for predicting biomedical events . . . 103
5.3 Materials and Methods . . . 106
5.3.1 Discretisation of continuous time series . . . 106
5.3.2 Discretised timestamped instance to sequential contrast patterns . . . 111
5.3.3 Illustrative examples of CHMM walking on sequential patterns . . . 117
5.4 Evaluation . . . 119
5.4.1 The septic shock prediction problem . . . 120
5.4.2 The MIMIC II database . . . 122
5.4.3 Selection of patients . . . 122

5.5 Prediction Results 123
5.5.1 Four data sets extracted from MIMIC-II 123
5.5.2 Cross-validation classification results on the four data sets 124
5.5.3 Predicting coupled discrete sequences using HMMs: An illustrative case study 125
5.5.4 Discussion 129
5.6 Conclusion 135

Chapter 6 Conclusions and Future Work 137
6.1 Conclusions 137
6.2 Future Work 140

Bibliography 142

List of Figures

3.1 Acute Hypotensive Episode over a time period exceeding 30 minutes, when MAP $\leqslant$ 60 mmHg 56

3.2 Observation and Target Windows with a Time Gap Interval . 57

3.3 Discretization by Symbolic Aggregate Approximation using 4 symbols 64

3.4 A Lexicographic Sequence Tree (LST) growing candidate sequences using 3 symbols as A, B, C 67

3.5 Effect of parameters L and G on the performance (A) For Event I, (B) For Event II 74

3.6 Inferring Visual Trends from Sequential Contrast Patterns Examples for AHE (A=1, B=2, C=3, D=4, E=5) 79

4.1 Transforming sequential patterns to binary or frequency based features. 90

4.2 The ICU Event Prediction Problem 93

5.1 A Lexicographic Sequence Tree (LST) growing candidate sequences using 3 symbols as X, Y, Z 111

5.2 Encoding and transformation of a data instance to an ordered sequence of patterns 113

5.3 Topology of a two-channel CHMM. 114

5.4 Encoding patient sequences using extracted patterns. P_j^i denotes a sequential pattern. Here, *i=1* indicates a single channel or variable. A patient MAP sequence such as $AACAABCBBC$ is converted to $P_1 - P_1 - P_3$. Finally, a new training set of pattern sequences is obtained. 117
5.5 State transition diagram with output emissions and their probabilities 118
5.6 A coupled HMM topology for 3 channels. Here, P_j^i denotes a sequential pattern. Here, i indicates a channel and j corresponds to a specific pattern-id for a variable. 120
5.7 Observation and Target Windows with a Time Gap Interval . 121

List of Tables

1.1 Transaction Data Table 2
1.2 Web Access logs 3

2.1 Transaction Data Example 18
2.2 Relational Attribute Data Example 18
2.3 Table 2.2 as a set of Transactions 19
2.4 Sequence Database 20

3.1 ICD-9 Classification of Hypotension 61
3.2 Checking gap constraint satisfaction of XY in XZXZY 68
3.3 Single Mode Classification Performance with 10 symbols . . . 71
3.4 Multi Mode Classification Performance with 15 symbols 72
3.5 Physionet 2009 AHE Test Prediction Classification Accuracies for events I and II given G=3 72
3.6 A Comparison of classification methods employed for the AHE prediction problem. Sequential patterns report comparable accuracies against existing methods 73
3.7 Representative Examples of Extracted AHE Sequential Patterns 77

4.1 Physionet AHE 2009 Test Prediction Results 94
4.2 MIMIC-II Hypotension Test Prediction Results 95
4.3 5-fold cross validated performances for Mortality Prediction . 96

5.1 A indicates the state transition function for discrete states S_1 and S_2 118

5.2 B denotes the emission probability distribution for 2 states and 4 pattern observations 118
5.3 A comparison of different models using 5-fold cross validation classification accuracy at $t_{gap} = 60$ mins and $t_{obs} = 60$ mins . . 125
5.4 A comparison of different models using 5-fold cross validation classification accuracy at $t_{gap} = 30$ mins and $t_{obs} = 60$ mins . . 126
5.5 A comparison of different models using 5-fold cross validation classification accuracy at $t_{gap} = 30$ mins and $t_{obs} = 90$ mins . . 126
5.6 A comparison of different models using 5-fold cross validation classification accuracy at $t_{gap} = 60$ mins and $t_{obs} = 90$ mins . 127
5.7 5-fold cross validation classification accuracy on CHMM and MCP-CHMM for 5 rounds of repeated re-sampling. g - gap interval size, o - observation window size 127
5.8 A multivariate (MAP, HR, RR) discrete patient sequence composed of an ordered series of contrast patterns 128
5.9 Visualizing contrast sequence patterns matching the three variables MAP, HR and RR 129
5.10 One way ANOVA Test on the 4 datasets (groups) corresponding to gap interval and observation window 131

List of Publications

Papers Published

Peer-reviewed Journals

- **Shameek Ghosh**, Jinyan Li, Longbing Cao, Kotagiri Ramamohanarao (2017). Septic shock prediction for ICU patients via coupled HMM walking on sequential contrast patterns. **Journal of Biomedical Informatics**, Elsevier, Volume 66, February 2017, Pages 19-31.

- **Shameek Ghosh**, Mengling Feng, Hung Nguyen, Jinyan Li (2016). Hypotension Risk Prediction via Sequential Contrast Patterns of ICU Blood Pressure. **IEEE Journal of Biomedical and Health Informatics**, Volume 20, Issue 5, 2016, pp. 1416-1426.

Peer-reviewed Conferences

- **Shameek Ghosh**, Hung Nguyen, Jinyan Li (2016). Predicting short-term ICU outcomes using a sequential contrast motif based classification framework. *in* Proceedings of IEEE Annual International Conference of the Engineering in Medicine and Biology Society (**EMBC-2016**), pp. 5612-5615. IEEE.

- **Shameek Ghosh**, Jinyan Li (2015), Using sequential patterns as features for classification models to make accurate predictions on ICU events. *in* Proceedings of IEEE Annual International Conference of the Engineering in Medicine and Biology Society (**EMBC-2015**), IEEE.

- **Shameek Ghosh**, Mengling Feng, Hung Nguyen, Jinyan Li (2014). Risk Prediction for Acute Hypotensive Patients by Using Gap Constrained Sequential Contrast Patterns. *in* Proceedings of the American Medical Informatics Association Annual Symposium (**AMIA-2014**), pp. 1748-1757.

- **Shameek Ghosh**, Mengling Feng, Hung Nguyen, Jinyan Li (2014), Predicting heart beats using co-occurring constrained sequential patterns. *in* Proceedings of Computing in Cardiology Conference (**CinC-2014**) pp. 265-268. IEEE.

Abstract

Data mining and knowledge discovery involves efficient search and discovery of patterns in data that are able to describe the underlying complex structure and properties of the corresponding system. To be of practical use, the discovered patterns need to be novel, informative and interpretable. Large-scale unstructured biomedical databases such as electronic health records (EHRs) tend to exacerbate the problem of discovering interesting and useful patterns. Typically, patients in intensive care units (ICUs) require constant monitoring of vital signs. To this purpose, signicant quantities of patient data, coupled with waveform signals are gathered from biosensors and clinical information systems. Subsequently, clinicians face an enormous challenge in the assimilation and interpretation of large volumes of unstructured, multidimensional, noisy and dynamically fluctuating patient data.

The availability of de-identified ICU datasets like the MIMIC-II (Multiparameter Intelligent Monitoring in Intensive Care) databases provide an opportunity to advance medical care, by benchmarking algorithms that capture subtle patterns associated with specific medical conditions. Such patterns are able to provide fresh insights into disease dynamics over long time scales.

In this research, we focus on the extraction of computational physiological markers, in the form of relevant medical episodes, event sequences and distinguishing sequential patterns. These interesting patterns known as sequential contrast patterns are combined with patient clinical features to develop powerful clinical prediction models. Later, the clinical models are

used to predict critical ICU events, pertaining to numerous forms of hemodynamic instabilities causing acute hypotension, multiple organ failures, and septic shock events. In the process, we employ novel sequential pattern mining methodologies for the structured analysis of large-scale ICU datasets. The reported algorithms use a discretised representation such as symbolic aggregate approximation for the analysis of physiological time series data. Thus, symbolic sequences are used to abstract physiological signals, facilitating the development of efficient sequential contrast mining algorithms to extract high risk patterns and then risk stratify patient populations, based on specic clinical inclusion criteria.

Chapter 2 thoroughly reviews the pattern mining research literature relating to frequent sequential patterns, emerging and contrast patterns, and temporal patterns along with their applications in clinical informatics.

In Chapter 3, we incorporate a contrast pattern mining algorithm to extract informative sequential contrast patterns from hemodynamic data, for the prediction of critical care events like Acute Hypotension Episodes (AHEs). The proposed technique extracts a set of distinguishing sequential patterns to predict the occurrence of an AHE in a future time window, following the passage of a user-defined gap interval. The method demonstrates that sequential contrast patterns are useful as potential physiological biomarkers for building optimal patient risk stratification systems and for further clinical investigation of interesting patterns in critical care patients.

Chapter 4 reports a generic two stage sequential patterns based classification framework, which is used to classify critical patient events including hypotension and patient mortality, using contrast patterns. Here, extracted sequential patterns undergo transformation to construct binary valued and frequency based feature vectors for developing critical care classification models.

Chapter 5 proposes a novel machine learning approach using sequential contrast patterns for the early prediction of septic shock. The approach combines highly informative sequential patterns extracted from multiple phys-

iological variables and captures the interactions among these patterns via Coupled Hidden Markov Models (CHMM). Our results demonstrate a strong competitive accuracy in the predictions, especially when the interactions between the multiple physiological variables are accounted for using multivariate coupled sequential models. The novelty of the approach stems from the integration of sequence-based physiological pattern markers with the sequential CHMM to learn dynamic physiological behavior as well as from the coupling of such patterns to build powerful risk stratification models for septic shock patients.

All of the described methods have been tested and bench-marked using numerous real world critical care datasets from the MIMIC-II database. The results from these experiments show that multivariate sequential contrast patterns based coupled models are highly effective and are able to improve the state-of-the-art in the design of patient risk prediction systems in critical care settings.