Mining of predictive patterns in Electronic health records data

Iyad Batal and Milos Hauskrecht
Department of Computer Science
University of Pittsburgh
milos@cs.pitt.edu

1 Introduction

The emergence of large-scale datasets in health care that record large amounts of information about the patients, their diseases and treatments and provide us with an opportunity to understand better the dynamics of the disease, efficacy of treatments, and various influences affecting the well-being of a patient. The development of computer methods and tools that would enable to analyze and utilize such data is badly needed.

The objective of this report is to briefly review data mining and machine learning methods we have developed since 2009 [7, 10, 6, 11, 8, 9, 12, 5, 4, 13] that aim to extract predictive patterns characterizing patient subgroups and their predictive differences in electronic health records (EHRs). Identification of such predictive patterns can be extremely useful for both knowledge discovery purposes and for feature engineering facilitating the construction of various predictive models from complex clinical time-series data.

The temporal pattern mining approach we have investigated and developed builds upon the temporal abstraction framework [19], that converts multivariate time-series data into sequences of discrete values, and combines their components using temporal logic operators [2] to form more complex temporally extended patterns. We have studied and made the following contributions to this line of work. First, we have proposed and developed a new *minimal predictive pattern mining approach* that lets us describe the predictive differences in the data by stratifying the data into subpopulations represented by patterns of different complexity, that let us better predict the patient outcomes. Second we have developed and tested two statistical criteria to filter random (or spurious) patterns that represent subpopulation with significantly different outcomes. Finally, we have developed and tested a new pattern mining heuristic that can be applied to event detection tasks. The heuristic searchers for minimum predictive patterns backwards in time, starting from the time of the event, such that more recent patterns with respect to the event are preferred and more likely included in the predictive pattern set constructed by the algorithm.

We have shown our approaches leads to a smaller number of predictive patterns

than other pattern-mining methods (that often retain many spurious patterns), while they preserve their combined prediction strength. Hence our methods enable a more compact description of predictive differences among patient groups. The majority of our methods were tested on postsurgical cardiac patient data (PCP dataset) extracted and applied in our previous work [16, 15, 18, 21, 14, 20].

2 Background

Frequent pattern mining is a very popular data mining technique to extract patterns from the data. Since their introduction in [1], frequent pattern and association rule mining have received a great deal of attention and have been successfully applied to various domains. Briefly, frequent pattern mining aims to identify all combinations of attribute-value pairs (defining patterns) that occur frequently in the data. Each such pattern is then characterized by a support that reflects the frequency of occurrence of the pattern in the dataset. Frequent patterns obey the *monotonicity* property: "if pattern P is not frequent, then all its super-patterns ($P'' \supset P$) are not frequent". This property implies that if P is frequent, then all its sub-patterns ($P' \subset P$) are also frequent. The basic algorithm for finding all frequent patterns, the apriori algorithm, searches systematically the space of patterns staring from more general patterns and continuing with more refined sub-patterns to identify those that surpass some predefined minimal support threshold.

The frequent pattern mining idea and algorithms for finding frequent patterns have been used or modified to identify other, more complex objects or relations in data. One of the most common extensions, the association rule mining framework, seeks to identify relationships among patterns, typically, a relation in between a pattern that consists of a logical combination of attribute-value pairs (antecedent of the rule) and a singleton pattern that consists of just one attribute-value pair (consequent of the rule). The association rule is then characterized by its support (how frequently the antecedent pattern occurs in data) and its precision (how often the consequent is associated with a satisfied antecedent).

In our work we are interested in applying pattern mining in the supervised (classification) setting where we have a specific target variable (outcome variable) and we want to identify patterns and sub-patterns (groups) in the EHR that are important for explaining and predicting this variable. An example of such patterns are: "a subpopulation of patients who received drug *a* has lower incidence of the recurrence of symptom b than the rest of the patients".

The current pattern mining algorithms are not fit well for analyzing clinical data. First the standard pattern-mining algorithms typically search exhaustively the feature space for all patterns with the minimum support threshold which is not feasible for high dimensional spaces we must work with in EHR. Second, the patterns these methods output are independent of each other and the output is a long list of association rules that is hard to read and comprehend. Third the feature space itself for EHR is very hard to define. The EHR data are temporal and are very different from other types of data used in frequent pattern mining framework. Briefly, each EHR consists of complex multivariate time series of clinical variables collected for a specific patient, such as

laboratory test results, medication orders, physiological parameters, past patient's diagnoses, surgical interventions and their outcomes. The times series can be of different length. Second, the temporal data in EHRs are acquired asynchronously, which means they are measured at different time moments and are irregularly sampled in time. This prevents from directly applying many standard time series algorithms to analyze them.

3 Methods

Our objective is to develop methods for finding predictive patterns representing subpopulations of patients with siginificantly different outcomes than the rest of the population. An example of a predictive pattern is: "a subpopulation of patients who received drug *a* has lower incidence of the recurrence of symptom b than the rest of the patients".

Identification of such predictive patterns is important for: (1) knowledge discovery and (2) feature engineering purposes. In knowledge discovery the goal is to identify and report such a pattern to human. This process may lead to the discovery of unexpected relationships not known prior to the analysis. In feature engineering, the pattern when it is present may lead to improved prediction of the outcome variable, e.g. "the recurrence of symptom b" from the previous example. In the following, we focus our discussion more on the utility of predictive temporal patterns for building and learning predictive classification models.

health record systems.

The main challenge for building classification models for EHR data is to define a good set of features that are able to represent well the temporal aspect of the data important for the prediction. We have studied two approaches to solve this challenge. First, we have build features by defining fixed feature mappings for each type of clinical variable in EHR and its time series [16, 22]. The limitation of this approach is that the mappings must be predefined by experts and are not built from data. Our second approach builds features corresponding to predictive temporal patterns automatically from data. We have developed the new pattern mining approach and published it in multiple papers [7, 10, 6, 11, 8, 9, 12, 5, 4, 13]. The approach relies on temporal abstractions and temporal pattern mining to extract a useful set of classification features.

Temporal abstraction features. To define 'flexible' features for classifying temporal data in EHRs we studied and implemented temporal abstractions methods that let us translate temporal data into sequences of temporal abstractions [19]. The temporal abstraction approach first converts time series for all clinical variables into sequences of discrete values and their time-intervals. Figure 1 illustrates the conversion of values for platelets lab using *value and trend abstractions*, where abstractions correspond to discrete categories such as high value of platelets or increasing platelets. The temporal abstractions can be then combined to form more complex temporal patterns using a set of temporal logic relations [2, 17] (e.g. descrease in platelets co-occuring with decrease in Hemoglobin, or, the administration of heparin followed by the drop in platelets). We can view these abstractions as definitions of temporal patterns describing the subpopulations of patients.

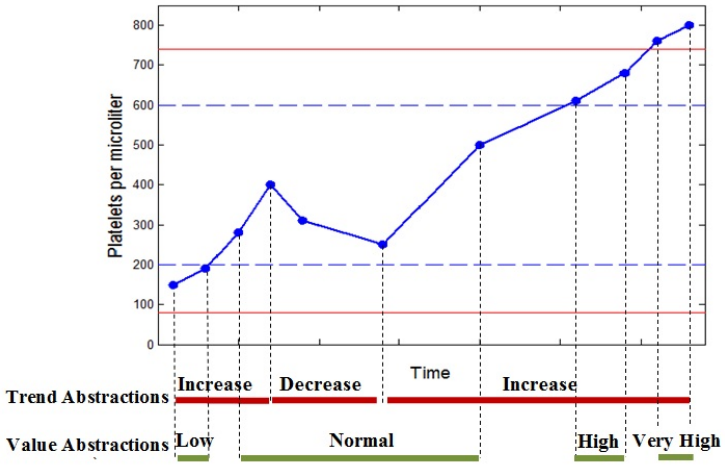

Figure 1: An example illustrating the trend and value abstractions. The value abstraction is defined using five discrete values (Very Low, Low, Normal, High, Very High), the trend abstraction using three values (Increase, Steady, Decrease).

Predictive pattern mining. The features based on temporal abstractions let us represent temporal patterns in time series data in the electronic health record. However, not all temporal pattern features are important for the classification task we want to solve. In particular we want to generate only features that help us to classify well the patient time series. To address this problem, our first approach [10, 11] adopted the frequent pattern mining paradigm that scanned groups and subgroups of data instances by gradually refining their temporal abstractions till they reached some minimum support threshold. The set of most discriminative patterns was then selected in the second pass. We found this approach to be inefficient whenever the number of temporal variables considered became large. First, the many patterns generated in the first phase were irrelevant for the prediction, so huge amount of time was spent on checking and scanning these patterns. Second, many predictive patterns selected were redundant or spurious in that they did not bring any (or very little) new information when compared to more general patterns that represented larger populations.

Minimal predictive pattern mining. To address these problems we have developed a new class of algorithms [8, 9] that automatically build, scan and mine only those temporal abstraction patterns that are important for the classification task represented by the target class variable. In addition, we proposed and implemented the minimal predictive pattern mining approach that eliminates the pattern redundancies by selecting only those discriminative temporal patterns that are significantly different (according to a confidence parameter α) from more general patterns in terms of their prediction strength. In [9] and later in [12] we showed that features generated by our minimal predictive pattern mining algorithm lead to a small number of patterns (features) that, when combined with an SVM classifier, lead to significant improvements in its classi-

fication performance. More specifically, each temporal pattern selected by our method was run on all data instances in the dataset, and used to generate either the value 1 if pattern was observed or 0 if it was not. This information became a feature the SVM classification model used to learn the classification model. We tested our approach on the problem of predicting patients who are at risk of developing heparin induced thrombocytopenia (HIT) [24, 23], an adverse clinical condition, that may lead to serious complications (thrombosis) and even death if it is not managed promptly. The results demonstrated the benefit of our approach in learning accurate classifiers, which was important for developing intelligent clinical monitoring systems.

Pattern significance and selection. In the minimal predictive pattern mining framework the significance of a candidate pattern is assessed in terms of a statistical score measuring the consistency of the pattern with its sub-patterns. We have developed and tested two scoring criteria for filtering the redundant patterns: the binomial score [12], and the Bayesian score [4]. To calculate the binomial score [12] we first estimate the probability of observing the outcome $\theta = P(Y = 1|G)$ for the pattern G using the maximum likelihood estimate. After that we check and calculate the chance (using the binomial distribution) that the examples in data satisfying a sub-pattern P are consistent with G in terms of outcomes $Y = 1$. The pattern that is significantly different (at some level) from G is preserved, otherwise it is excluded. In [4] we proposed a more robust Bayesian test [4] that lets us assess the chance of a more predictive sub-pattern being generated by chance. Let M_1 be the model that assumes that all instances of G have the same probability for outcome $Y = 1$, even though we are uncertain what that probability is. We denote this probability by θ. Now us now assume a model M_2 for which the probability of $Y = 1$ in the subpopulation of G represented by pattern P is θ_1 and the probability of $Y = 1$ for the complement of P on G, P, is θ_0. Let us also assume that $\theta_1 > \theta_0$ for M_2. To represent our uncertainty about parameters θ, θ_1 and θ_0 and their distribution we assign them the noninformative Beta distribution prior. The Bayesian score, reflecting the usefulness of the pattern P for defining M_2, is the difference in the marginal likelihoods $Bscore(P) = P(data|M_2) - P(data|M_1)$.

Recent predictive patterns. The space of possible temporal patterns one can define with the temporal abstractions is enormous. The key challenge is to find ways of reducing the complexity of this space as much as possible, hence improving the efficiency of the mining algorithms. To address this concern we proposed and started to study a new approach that builds predictive patterns for monitoring and event detection problem using the 'recent predictive pattern' heuristic [5], which captures the intuition that most recent information related to the clinical variable is likely the most important for the future prediction. We succesfuly tested the initial version of this approach approach by predicting: (1) the patients who are at risk of developing heparin induced thrombocytopenia (HIT) [3], and (2) complications (diagnosis of secondary diseases) for diabetes patients [5].

4 Conclusions

We have reviewed our recent effort and advances we have made in developing predictive pattern mining framework based on temporal abstractions. However, many critical issues remain open and a number of improvements of the framework are possible. First, the most critical open problem is to find ways for reducing the space in which the predictive patterns are searched. Efficient heuristics for restricting clinical variables or values the variables may take when building and searching more complex patterns are neccessary and need to be designed. Second, the patterns/features based on temporal abstractions may be used either for building better classification models or for the purpose of knowledge discovery when subgroups. While the benefit of the patterns for classification can be judged by the analysis of models built using these features, their utility for knowledge discovery needs to be further investigated and would require a carefully designed evaluation studies with human evaluators.

5 References

References

[1] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the international conference on Management of data (SIGMOD)*, 1993.

[2] F. Allen. Towards a general theory of action and time. *Artificial Intelligence, 23:123-154*, 1984.

[3] Iyad Batal. *Mining predictive patterns and extension to multivariate temporal data*. PhD thesis, University of Pittsburgh, 2012.

[4] Iyad Batal, Gregory Cooper, and Milos Hauskrecht. A bayesian scoring technique for mining predictive and non-spurious rules. In *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 260–276. Springer Berlin Heidelberg, 2012.

[5] Iyad Batal, Dmitriy Fradkin, James Harrison, Fabian Moerchen, and Milos Hauskrecht. Mining recent temporal patterns for event detection in multivariate time series data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 280–288, New York, NY, USA, 2012. ACM.

[6] Iyad Batal and Milos Hauskrecht. Boosting knn text classification accuracy by using supervised term weighting schemes. In *Proceeding of the 18th ACM conference on Information and knowledge management*, pages 2041–2044. ACM, 2009.

[7] Iyad Batal and Milos Hauskrecht. A supervised time series feature extraction technique using dct and dwt. In *International Conference on Machine Learning and Applications (ICMLA)*, 2009.

[8] Iyad Batal and Milos Hauskrecht. A concise representation of association rules using minimal predictive rules. In *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)*, 2010.

[9] Iyad Batal and Milos Hauskrecht. Constructing classification features using minimal predictive patterns. In *Proceedings of the international conference on Information and knowledge management (CIKM)*, 2010.

[10] Iyad Batal, Lucia Sacchi, Riccardo Bellazzi, and Milos Hauskrecht. Multivariate time series classification with temporal abstractions. In *Proceedings of the Florida Artificial Intelligence Research Society (FLAIRS)*, 2009.

[11] Iyad Batal, Lucia Sacchi, Riccardo Bellazzi, and Milos Hauskrecht. A temporal abstraction framework for classifying clinical temporal data. *AMIA Annual Symposium Proceedings*, 2009:29, 2009.

[12] Iyad Batal, Hamed Valizadegan, Gregory F. Cooper, and Milos Hauskrecht. A pattern mining approach for classifying multivariate temporal data. In *Proceedings of the IEEE international conference on bioinformatics and biomedicine (BIBM)*, 2011.

[13] Iyad Batal, Hamed Valizadegan, Gregory F Cooper, and Milos Hauskrecht. A temporal pattern mining approach for classifying electronic health record data. *Transactions on Intelligent Systems and Technology*, 4:4, 2013.

[14] M. Hauskrecht, M. Valko, B. Kveton, S. Visweswaran, and G.F. Cooper. Evidence-based anomaly detection in clinical domains. In *AMIA Annual Symposium Proceedings*, pages 319 – 324, 2007.

[15] Milos Hauskrecht, Iyad Batal, Michal Valko, Shyam Visweswaram, Gregory Cooper, and Gilles Clermont. Outlier detection for patient monitoring and alerting. *Journal of Biomedical Informatics*, 46:47–55, 2013.

[16] Milos Hauskrecht, Michal Valko, Iyad Batal, Gilles Clermont, Shyam Visweswaram, and Gregory Cooper. Conditional outlier detection for clinical alerting. In *Proceedings of the American Medical Informatics Association (AMIA)*, 2010.

[17] Frank Höppner. *Knowledge Discovery from Sequential Data*. PhD thesis, Technical University Braunschweig, Germany, 2003.

[18] Branislav Kveton and Milos Hauskrecht. Solving factored mdps with exponential-family transition models. In *Proceedings of the 16th International Conference on Automated Planning and Scheduling*, pages 114–120, 2006.

[19] Y. Shahar. A Framework for Knowledge-Based Temporal Abstraction. *Artificial Intelligence, 90:79-133*, 1997.

[20] Hamed Valizadegan, Quang Nguyen, and Milos Hauskrecht. Learning classification models from multiple experts. *Journal of biomedical informatics*, 2013.

[21] M. Valko and M. Hauskrecht. Feature importance analysis for patient management decisions. In *13th International Congress on Medical Informatics*, pages 861 – 865. NIH Public Access, 2010.

[22] Michal Valko and Milos Hauskrecht. Feature importance analysis for patient management decisions. In *Proceedings of medical informatics (MedInfo)*, 2010.

[23] TE. Warkentin. Heparin-induced thrombocytopenia: pathogenesis and management. *Br J Haematology*, pages 535 – 555, 2003.

[24] TE. Warkentin, JI. Sheppard, and P. Horsewood. Impact of the patient population on the risk for heparin-induced thrombocytopenia. *Blood*, pages 1703 – 1708, 2000.