Journal of Theoretical and Applied Information Technology
10th January 2014. Vol. 59 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# SIGNATURE BASED MINING FRAMEWORK FOR EVENT SEQUENCES AND ITS APPLICATIONS IN HEALTHCARE DATA USING LANGUAGE MODEL

**[1]D.VETRITHANGAM, [2]Dr.N.UMA MAHESHWARI [3]Dr.R.VENKATESH**

[1]Assistant Prof., Department of Computer Science and Engineering,
RVS College of Engineering and Technology, Dindigul,Tamilnadu, India
[2] Professor., Department of Computer Science and Engineering,
PSNA College of Engineering and Technology,Dindigul , Tamilnadu, India
[3] Professor., Department of Information Technology,
PSNA College of Engineering and Technology,Dindigul , Tamilnadu, India
E-mail: [1]vetrigold@gmail.com , [2]numamahi@gmail.com

**ABSTRACT**

Temporal event signature mining for knowledge discovery is a difficult problem. In this paper a framework is designed to know a temporal knowledge about the large scales signature mining of longitudinal heterogeneous event data. This framework mainly deals with the mining of high order latent event structure and its relationship within single and multiple event sequences. Here, the heterogeneous event sequences maps to geometric image by encoding structured into spatial temporal shape process. A Probabilistic language modeling is used to extract high order events from large scale data. Also, presents a doubly constrained conventional sparse coding to learn interpretable and shift invariant latent temporal event signature. This can be shown by the sparsity in the data and latent factor model on the β-divergence. An optimization scheme is also used to perform large scale incremental learning of group specific temporal event signatures.

**Keywords:** *Data mining(DM), Signature mining(SM), knowledge representation(KR), Probabilistic Model(PM) , Cluster (C)*

## 1. INTRODUCTION

Finding latent temporal signatures is important in many domains as they encode temporal concepts such as event trends, episodes, cycles, and abnormalities. For example, in the medical domain latent event signatures facilitate decision support for patient diagnosis, prognosis, and management. In the surveillance domain temporal event signatures aid in detection of suspicious events at specific locations. For instance, in the medical domain a patient is considered as an event entity, where a visit to the doctor's office is considered as events.

Temporal event signature mining for knowledge discovery is a difficult problem. The vast amounts of complex event data pose challenges not only for humans, but also for data and information analysis by machines. Two fundamental questions in addressing this challenge are: What is an appropriate knowledge representation for mining longitudinal event data and how can we learn such representation from large complex datasets? An event knowledge representation (EKR) is commensurate with human capabilities so complex event data can quickly be absorbed, understood, and efficiently transformed into actionable knowledge. In this regard, several problems need to be addressed:

The EKR should handle the time-invariant representation of multiple event entities as two event entities can be considered similar if they contain the same temporal signatures at different time intervals or locations, EKR should be flexible to jointly represent different types of event structure such as single multivariate events and event intervals to allow a rich representation of complex event relationships, EKR should be scalable to support analysis and inference on large-scale databases, and this paper proposes a novel temporal event matrix representation and learning framework to perform temporal signature mining for large-scale longitudinal and heterogeneous event data. Basically, our TEMR framework represents the event data as a spatial-temporal matrix, where one dimension of the matrix corresponds to the type of the events and the other dimension represents the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

time information. In this case, if event i happened at time j with value k, then the (i, j) th element of the matrix is k. This is a very flexible and intuitive framework for encoding the temporal knowledge information contained in the event sequences. Fig. 1 illustrates a simple example on represent-ing the longitudinal medical record of a diabetic patient over one year using our TEMR approach, where the vertical axis corresponds to the different events the horizontal axis represents the time information associated with these events. There is a dot in the matrix if the corresponding event happened at the corresponding time. Because of the analogy between matrix and image, TEMR offers a flexible and intuitive way of encoding comprehensive temporal knowledge, including event order-ing, duration, and heterogeneity. With this representation, we develop a matrix approximation-based technology to detect the hidden signatures from the event sequences. We prove theoretically the convergence of the proposed algorithm. To improve the scalability of the proposed approach, we further developed an online updating technology. Finally, the effectiveness of the proposed algorithm is validated on a real-world healthcare dataset. It is worthwhile to outline the advantages of the proposed approach.

First, on the knowledge representation level, TEMR provides a visual matrix-based representation of complicated event data composed of different types of events as well as event intervals, which supports the joint representation of both continuous and discrete valued data.

Second, on the algorithmic level, proposing a doubly sparse convolutional matrix approximation-based formulation for detecting the latent signatures contained in the datasets. Moreover, we derive a multiplicative updates procedure to solve the problem and proved theoretically its convergence. We further propose a novel stochastic optimization scheme for large-scale longitudinal event

## 2. DATA MINING IN MEDICINE

### 2.1 Data Mining in the Health Sector

Today, the size of the population, the amount of electronic data gathered, along with globalization and the speed of disease outbreaks make it almost impossible to accomplish what the pioneers did. This is where data mining becomes useful to healthcare. It has been slowly but increasingly applied to tackle various problems of knowledge discovery in the health sector

Data mining and its application to medicine and public health is a relatively young field of study. In 2003, Wilson et al began to scan cases where KDD and data mining techniques were applied in health databases. They found confusion in the field regarding what constituted data mining. “Some authors refer to data mining as the process of acquiring information, whereas others refer to data mining as utilization of statistical techniques within the knowledge discovery process.” (Wilson et al. 2003) Because of misconceptions still going on in the medical community about what data mining comprises, let us first define what we mean by it. The generally accepted definition of data mining today is the set of procedures and techniques for discovering and describing patterns and trends in data (Witten and Frank 2005).

### 2.2 The Importance and Uses of Data Mining in Medicine and Public Health

Despite the differences and clashes in approaches, the health sector has more need for data mining today. There are several arguments that could be advanced to support the use of data mining in the health sector, covering not just concerns of public health but also the private health sector (which, in fact, as can be shown later, are also stakeholders in public health).Data overload there is a wealth of knowledge to be gained from computerized health records. Yet the overwhelming bulk of data stored in these databases makes it extremely difficult, if not impossible, for humans to sift through it and discover knowledge In fact, some experts believe that medical breakthroughs have slowed down, attributing this to the prohibitive scale and complexity of present-day medical information. Computers and data mining are best-suited for this purpose.

### 2.3 Evidence-Based Medicine and Prevention of Hospital Errors

When medical institutions apply data mining on their existing data, they can discover new, useful and potentially life-saving knowledge that otherwise would have remained inert in their databases. For instance, an ongoing study on hospitals and safety found that about 87% of hospital deaths in the United States could have been prevented, had hospital staff (including doctors) been more careful in avoiding errors (HealthGrades Hospitals Study 2007). By mining hospital records, such safety issues could be flagged and addressed by hospital management and government regulators.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

### 2.4 Policy-Making in Public Health

Lavrac et al. (2007) combined GIS and data mining using among others, Weka with J48 (free, open source, Java-based data mining tools), to analyze similarities between community health centers in Slovenia. Using data mining, they were able to discover patterns among health centers that led to policy recommendations to their Institute of Public Health. They concluded that "data mining and decision support methods, including novel visualization methods, can lead to better performance in decision-making." The preceding factors remind us of an incident in the Philippines at the Rizal Medical Center in Pasig City in October 2006. Failing to implement strict sanitation and sterilization measures the hospital contributed to the death of several new-born babies due to neonatal sepsis (bacterial infection). No one really knew what was going on until the deaths became more frequent. Upon examining hospital records, the Department ofHealth (DOH) found that 12 out of 28 babies born on October 4, for example, died of sepsis (Tandoc 2006). With an integrated database and the application of data mining the DOH could detect such unusual events and curtail them before they worsen.

### 2.5 More Value for Money and Cost Savings

Data mining allows organizations and institutions to get more out of existing data at minimal extra cost. KDD and data mining have been applied to discover fraud in credit cards and insurance claims (Kou et al. 2004). By extension, these techniques could also be used to detect anomalous patterns in health insurance claims, particularly those operated by Phil Health, the national healthcare insurance system for the Philippines.

### 2.6 Early Detection and/or Prevention of Diseases

Cheng, et al cited the use of classification algorithms to help in the early detection of heart disease, a major public health concern all over the world. Cao et al (2008) described the use of data mining as a tool to aid in monitoring trends in the clinical trials of cancer vaccines. By using data mining and visualization, medical experts could find patterns and anomalies better than just looking at a set of tabulated data.

### 2.7 Early Detection and Management of Pandemic Diseases and Public Health Policy Formulation

Health experts have also begun to look at how to apply data mining for early detection and management of pandemics. Kellogg et al. (2006) outlined techniques combining spatial modeling, simulation and spatial data mining to find interesting characteristics of disease outbreak. The analysis that resulted from data mining in the simulated environment could then be used towards more informed policy-making to detect and manage disease outbreaks. Wong et al. (2005) introduced WSARE, an algorithm to detect outbreaks in their early stages. WSARE, which is short for "What's Strange about Recent Events", is based on association rules and Bayesian networks. Applying WSARE on simulation models have been claimed to result to relatively accurate predictions of simulated disease outbreaks. Of course, these sorts of claims always come with warnings to take precaution when applying these models in real life.

### 2.8 Non-Invasive Diagnosis and Decision Support

Some diagnostic and laboratory procedures are invasive, costly and painful to patients. An example of this is conducting a biopsy in women to detect cervical cancer. Thangavel et al (2006) used the K-means clustering algorithm to analyze cervical cancer patients and found that clustering found better predictive results than existing medical opinion. They found a set of interesting attributes that could be used by doctors as additional support on whether or not to recommend a biopsy for a patient suspected of having the cervical cancer.

## 3. LANGUAGE MODELING OF EVENT SEQUENCES

A probabilistic model of the event sequences is developed in this corpus. In selecting a type of model to use, our goals were to find a model that would allow for a good fit to the data, but also one that was computationally feasible. we chose the SRI Language Modeling toolkit (SRILM) to build a model for the sequence of events in the form of a traditional language model (Stolcke,2002). Language modeling has proven that it is able to model sequential data like sequences of words, tags, and phonemes very well in other natural language processing tasks. There is an intuition for using a language model for sequences of events.

Language modeling also has a number of characteristics that make it well suited for modeling event sequences in our corpus. Tools like the SRI Language Modeling toolkit are extremely fast, allowing us to analyze significant portions of our corpus. To use this language model, one can estimate the probability of a sequence of events by

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

calculating n-gram probabilities. These probabilities can then be used in many different story understanding applications, such as event ordering, event prediction, story classification, story similarity evaluation and automated story generation.

## 4. TEMPORAL EVENT SIGNATURE MINING

Introducing the details of how to detect temporal event signatures with our TEMR representation. First, by introducing some preliminaries.

### 4.1 Preliminaries

Consider a event matrix X €$IR^{n*t}$, where n is the number of different event types, t is the length of the event sequence. As mentioned in Section 3.2, we assume X is the superposition of the one-side convolution of a set of hidden patterns F ={$F^{®}$}$^{R}_{r=1}$ across the time axis. We define the one-side convolutional operator * as follows

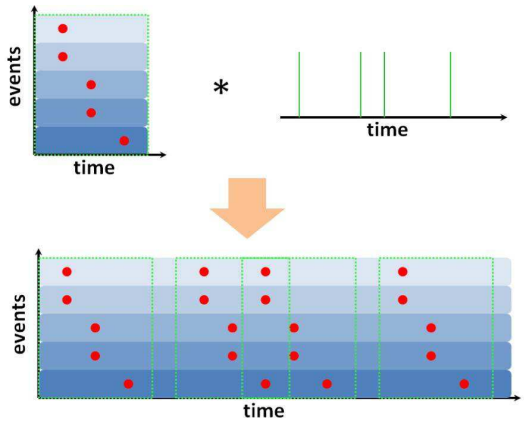


*Fig 1 A Graphical Illustration of One Side Convolution*

The top left figure shows the temporal signatures, and the top right figure is the time axis, where we use green bars to represent the position where the pattern appears. The bottom figure is the one-side convolution result, where each dotted line rectangle corresponds to a pattern.

Thus there is no convolution on the vertical axis. Fig. 1 gives us an intuitive graphical illustration of the procedure of one-side convolution, where the bottom image is obtained through the one-side convolution of such signature on top-left and the time vector on top-right.

### 4.2 Mining Signatures from Multiple Event Sequences

In many real-world scenarios, not only interested in discovering the signatures within a single event sequence, but also in detecting signatures from multiple event sequences. For example, in the medical domain, the event sequence of a single patient is usually very sparse. In this case, it makes more sense to detect signatures from a group of patients with similar disease conditions rather than a single patient.

More formally, consider the case where the event matrices are composed of n event sequences. Using X =[$X_1$; $X_2$; . . . ; $X_n$] to represent the event sequence group, with $X_l$ representing the lth event sequence in this group. In the following extending our one-side convolutional NMF to the scenarios. Still denoting the latent event signature set as histogram shape. The majority of patients in cluster IV exhibits a higher DCSI score and thus has higher risk of hospitalization and mortality. Clusters II and III show similar shapes of the overall histogram, indicating that the learned patterns within these patient groups mainly consist of common temporal signatures that are not indicative of disease severity. The longer right tail of the histogram can be explained by the rarity of patients who have a very high DCSI score. We note that one can go back to the individual patterns to investigate what kind of care the patients received.

F = {$F^{®}$}$^{R}_{r=1}$.

## 5. TEMPORAL SIGNATURE GROUPS VERSUS DIABETIC DISEASE

Generating a histogram that capture the patient distribution in each cluster. By performing visual examination of the patient distribution based on their severity level to look for group specific differences.

Fig. 2 shows an example of a four cluster partitioning of a random subset of our diabetic patient population. One can infer that the identified patterns in cluster IV mostly occur in groups of patients with a high DCSI score. Taking a closer look to cluster IV one can see the low number of patients with a low severity score (i.e., 1) in contrast to the overall

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

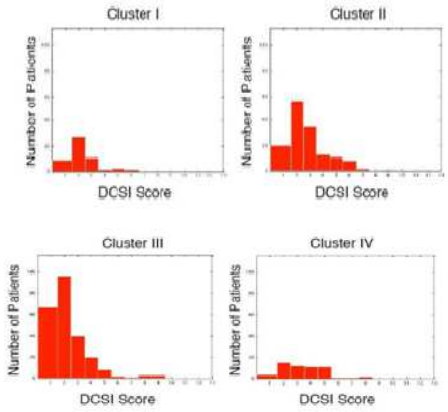


*Fig.2 Temporal Signature Groups Versus Diabetic Disease Severity Level.*

## 6. CONCLUSION

In this paper, we have presented a novel temporal event matrix representation and learning framework in conjunction with an in-depth validation on both synthetic and real world datasets. Our approach is to develop a means of extracting event sequences in health care data and apply this technology to an extremely large scale data, and use language modeling techniques to build a probabilistic model. The framework has wide applicability to a variety of data and application domains that involve large-scale longitudinal event data. We have demonstrated that our proposed framework is able to cope with the double sparsity problem and that the induced double sparsity constraint on the _-divergence enables automatic relevance determination for solving the optimal rank selection problem via an over complete sparse latent factor model. Further, the framework is able to learn shift invariant high-order latent event patterns in large-scale data. We empirically showed that our stochastic optimization scheme converges to a fixed point and we have demonstrated that our framework can learn the latent event patterns within a group. Future work will be devoted to a thorough clinical assessment for visual interactive knowledge discovery in large electronic health record databases.

## REFRENCES:

[1] B. Cao, D. Shen, J.T. Sun, X. Wang, Q. Yang, and Z. Chen, "Detect and Track Latent Factors with Online Nonnegative Matrix Factorization," Proc. 20th Int'l Joint Conf. Artificial Intelligence, 2689-2694, 2007.

[2] F.R.K. Chung, Spectral Graph Theory. Am. Math. Soc., 1997.

[3] C. Ding, T. Li, and M.I. Jordan, "Convex and Semi-Nonnegative Matrix Factorizations," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 32, no. 1, pp. 45-55, Jan. 2010.

[4] M. Dong, "A Tutorial on Nonlinear Time-Series Data Mining in Engineering Asset Health and Reliability Prediction: Concepts, Models, and Algorithms," Math. Problems in Eng., vol. 2010, pp. 1-23, 2010.

[5] C. Fe´votte and J. Idier, Algorithms for Nonnegative Matrix Factorization with the Beta-Divergence, arXiv:1010.1763, 2010.

[6] W. Fei, L. Ping, and K. Christian, "Online Nonnegative Matrix Factorization for Document Clustering," Proc. 11th SIAM Int'l Conf. Data Mining, 2011.

[7] D.D. Lee and H.S. Seung, "Learning the Parts of Objects by Non-Negative Matrix Factorization," Nature, vol. 401, no. 6755, pp. 788-91, 1999.

[9] P.O. Hoyer, "Non-Negative Matrix Factorization with Sparseness Constraints," J. Machine Learning Research, vol. 5, pp. 1457-1469, 2004.

[10] P.O. Hoyer, "Non-Negative Sparse Coding," Proc. 12th IEEE Workshop Neural Networks for Signal Processing, 2002.

[11] Mehdi Manshadi1, Reid Swanson2, and Andrew S. Gordon2 "Learning a Probabilistic Model of Event Sequences From Internet Weblog Stories

[12] Y.R. Ramesh Kumar and P.A. Govardhan, "Stock Market Predictions—Integrating User Perception for Extracting Better Prediction a Framework," Int'l J. Eng. Science, vol. 2, no. 7, 3305-3310, 2010.

[12] Stolcke, A. (2002) SRILM: An Extensible Language Modeling Toolkit, International Conference on Spoken Language Processing, Denver, Colorado.