Volume 3, Issue 12, December 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

# A Review on Mining Signatures from Event Sequences and Visual Interactive Knowledge Discovery in Large Electronic Health Record Database

**S.A.Sarwade**
*M.E. CSE 2nd Year*
*SSGBCOET,Bhusawal ,NMU,(M.S.),India*

**R.K.Makhijani**
*Associate Professor, CSE Dept.*
*SSGBCOET,Bhusawal NMU,(M.S.),India*

***Abstract---- Standardization and wider use of Electronic Health records (EHR) creates opportunities for better understanding patterns of illness and care within and across medical systems. In the healthcare systems, hidden event signatures allow taking decision for patient's diagnosis, prognosis, and management. Temporal history of event codes embedded in patients' records, investigates frequently occurring sequences of event codes across patients. There is a framework that enables the representation, extraction, and mining of high order latent event structure and relationships within single and multiple event sequences by mapping the heterogeneous event sequences to a geometric image by encoding events as a structured spatial-temporal shape process. Over the last decade, so many information visualization techniques have been developed to support the exploration of large data sets. There are various interactive visual data mining tools available for visual data analysis. It is possible to perform clinical assessment for visual interactive knowledge discovery in large electronic health record databases. In this paper, we proposed that it is possible to develop a tool for data visualization for interactive knowledge discovery. Data Visualization is very helpful for analysts to visually discover different kinds of patterns such as clusters, relationships and associations.***

***Keywords— Temporal signature mining, sparse coding, nonnegative matrix factorization, beta-divergence, Visual Analytics, Information visualization.***

## I. INTRODUCTION

In EHR data, each record (data instance) consists of multiple time series of clinical variables collected for a various patients, such as results of tests in laboratory and medication orders. The record may also provide information about patient's diseases and adverse medical events over time. Finding latent temporal signatures is important in many domains as they encode temporal concepts such as event trends, episodes, cycles, and abnormalities. Temporal data mining is concerned with data mining of large sequential data sets. By sequential data, we mean data that is ordered with respect to some index. For example, time series constitute a popular class of sequential data, in which records are arranged by time. Sequential data could be text, gene sequences, amino acid sequences, and moves in a puzzles or chess game. The ordering among the records is very important for the data description/modelling. Time series analysis has a long history. Techniques for statistical modelling and spectral analysis of real or complex-valued time series have been in use for more than fifty years. Temporal data mining methods must be capable of analysing data sets that are prohibitively large for conventional time series modelling techniques to handle efficiently. Temporal event signature mining for knowledge discovery is a difficult problem. Due to vast amounts of complex event data it is challenging for humans and also for data and information analysis by machines. An appropriate knowledge representation for mining longitudinal event data is important. This paper gives possibility is to provide interactive and user friendly representation of Knowledge and data with Visual Data Analytics.

## II. RELATED WORK

The important part of temporal data mining is its representation. An event knowledge representation (EKR) should be commensurate with human capabilities so complex event data can quickly be catch up , understood, and converted into actionable knowledge. Temporal data may be continuous or discrete. The popular approach is to transform the continuous time series data into discrete representations in the form of symbols such as string, nominal, categorical, and item sets for knowledge representation of continuous time data. For example, Lin et al. [13] summarized existing time series representations as data adaptive, such as Piecewise Linear Approximation (PLA), Adaptive Piecewise Constant Approximation (APCA), the Singular Value Decomposition (SVD), and Symbolic Aggregate approximation (SAX), and non-data adaptive, such as the standard Discrete Fourier transform (DFT), Discrete Wavelet Transform (DWT), and Piecewise Aggregate Approximation (PAA).For knowledge representation of discrete time series data, Moerchen et al. [15], [17], [16] proposed a novel Time Series Knowledge Representation (TSKR) as a pattern language (grammar) for temporal knowledge discovery from multivariate time series and symbolic data, where the temporal knowledge is represented in the form of symbolic languages and grammars that have been formulated as a

 

means to perform intelligent reasoning and inference from time-dependent event sequences. The TEMR framework in [1] provides another alternative way to represent the temporal knowledge contained in discrete data. As compared to symbolic and grammar-based representations, this approach is more intuitive and easy to understand. The relationships among all different types of events can clearly be observed using TEMR. It is very necessary to detect latent event signatures which are closely related to Nonnegative Matrix Factorization techniques. NMF is a very popular method which is used to extract the latent factors from nonnegative data matrix. Authors put forward an online NMF (ONMF) algorithm to detect latent factors and track their evolution while the data evolve [2].

Authors in [6] shows how to merge the concepts of non-negative factorization with conditions of sparsity. The result is a multiplicative algorithm which is comparable to standard NMF. Multiplicative algorithm can be used to obtain sensible solutions in the over complete cases. Study of online NMF (ONMF) algorithm to efficiently handle very large-scale and/or streaming datasets was proposed by authors in [7].Unlike conventional NMF solutions which require the entire data matrix to reside in the memory, ONMF algorithm proceeds with one chunk of data points at a time. It presents the experiments with one-pass and multi-pass ONMF on real datasets.Algorithms for nonnegative matrix factorization (NMF) with the β-divergence (β-NMF) are described in [8]. The β-divergence is a family of cost functions parametrized by a single shape parameter β that takes the Kullback-Leibler divergence, Euclidean distance, and the Itakura-Saito divergence as special cases (β = 2, 1, 0 respectively). Sparse NMF by adding a sparsity was introduced by Hoyer [9], [10] and Eggert [6] inducing regularizer to the standard NMF objective. This sparsity regularization further improves the model interpretability for efficient data representation. Convolutional NMF (cNMF) models have been proposed in Smaragdis [21] and O'Grady and Pearlmutter [18] to extract the latent sound objects from acoustic signals.

There is a large number of information visualization techniques which have been developed over the last decade to support the exploration of large data sets. In visual data exploration, the user is directly involved in the data mining process. Daniel A. Keim [22] propose a classification of information visualization and visual data mining techniques which is based on the data type to be visualized and the technique of visualization , interaction and distortion.

## III. EXISTING SYSTEM

Existing system [1] uses a novel Temporal Event Matrix Representation (TEMR) and learning framework to perform temporal signature mining for large-scale longitudinal and heterogeneous event data. TEMR framework represents the event data in the form of matrix, in which where one dimension corresponds to the type of the events and the other dimension represents the time information. If event 'i' happened at time 'j' with value 'm', then the (i, j)th element of the matrix is 'm'. This is a very flexible and intuitive framework for encoding the temporal knowledge information contained in the event sequences.

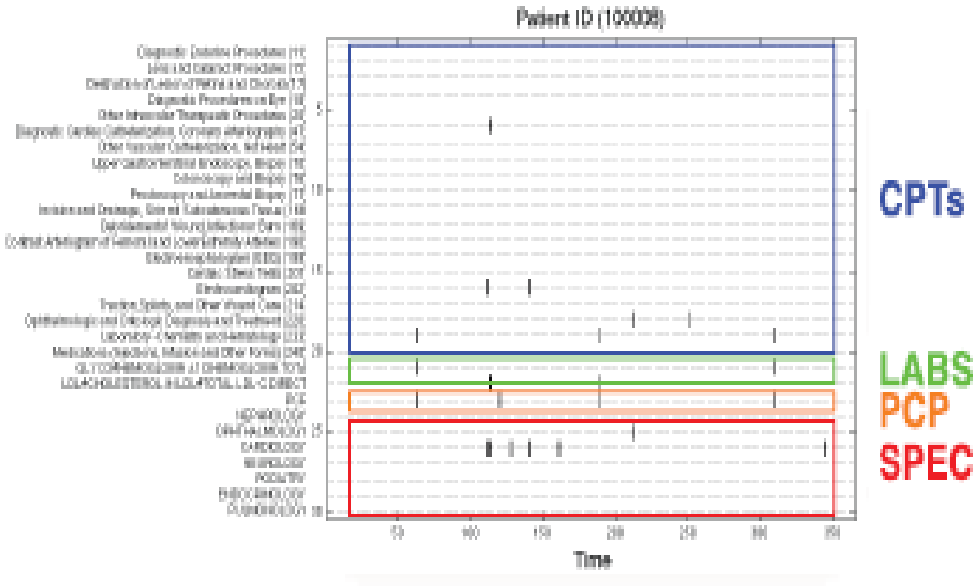


Fig 1. An example of a diabetic patient's electronic record over one year [1].

Figure.1 illustrates a simple example on representing the longitudinal medical record of a diabetic patient over one year using TEMR approach, in which the vertical axis corresponds to the different events such as primary care procedures, physician visits, lab tests, and specialist visits and the horizontal axis represents the time information associated with these events. A dot is used in the matrix for corresponding event happened at the corresponding time. There is analogy between matrix and image. TEMR gives a flexible and intuitive way of encoding comprehensive temporal knowledge. It contains event ordering, duration, and heterogeneity. Authors [1] developed a matrix approximation-based technology to detect the hidden signatures from the event sequences and developed an online updating technology. This enables the representation, extraction, and mining of high order latent event structure and relationships within single and multiple event sequences. The knowledge representation maps the heterogeneous event sequences to a geometric image by encoding events as a structured spatial-temporal shape process.

 

## IV. PROPOSED SYSTEM

The main objective proposed system is to provide interactive and user friendly representation of Knowledge and data with Visual Data Analytics. Visual data mining techniques are tightly integrated with the systems used to manage the vast amounts of relational and semi structured information, including database management and data warehouse systems. The final goal is to bring the power of visualization technology to every desktop to allow accurate, faster, and more intuitive exploration of very large data resources.

Visual Analytics [22] often comprises the usage of multiple views, which requires a well-designed and intuitive user interface taking into consideration the display and arrangement of the visualization and allow the user to interactively parameterize views. Visual analytics is used for the analysis of vast amounts of data to identify and visually distill the most valuable and relevant information content. In this system, automated analysis techniques are combined with the interactive visualizations. Due to this, it becomes very easy to understand and to make decision from given very large and complex data sets. Visual analytics creates tools and techniques which make possible for people to combine information and derive meaningful results from large, changing, indeterminate, and often inconsistent data, find out the expected and unexpected things. It also provides defensible, timely, and easily understandable assessments and effectively communicates assessment for action.

### *A. Visual Data Exploration*

Data Exploration means finding the valuable information hidden in like a drop in the ocean when dealing with data sets containing millions of data items. The aim of Visual data exploration is to integrate human in finding information from large, applying perceptual abilities to the large data sets available in computer systems. Visual data exploration presents the data in some visual form, and allows the human to get insight into the data, draw the conclusions, and allow interacting directly with the data. Visual data mining techniques have proven to be of high value in exploratory data analysis and they also have a high potential for exploring large databases. Visual data exploration can easily deal with highly inhomogeneous as well as noisy data. Visual data exploration is very intuitive and requires no understanding of complex mathematical or statistical algorithms or parameters.

### *B. The Visual Analytics Process*

Visual Analytics process let the user enter into a loop where data can be interactively manipulated to help gain insight both on the data and the representation itself. The sense-making loop structures of the whole knowledge discovery process are supported through Visual Analytics. The process then enters a loop where the user can gain knowledge on the data, ideally driving the system toward more focused and more adequate analytical techniques. The user will gain a better understanding of the visualization itself commanding for different views helping him or her to go beyond the visual and ultimately confirm hypotheses built from previous iterations [22] as shown in Figure 2.

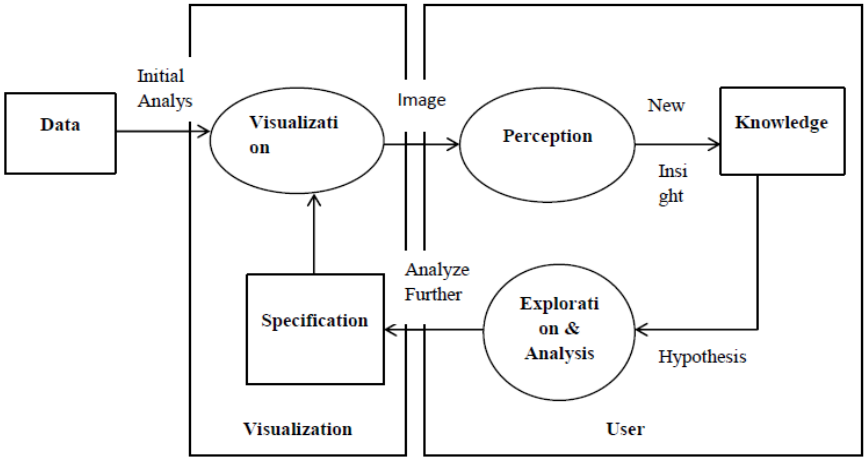


Fig. 2 Visual Analytics based on the simple model of Visualization

Standard data analysis systems provide a wide range of interactive visualization techniques and data views. There are number of tool and software available for visualization. In the proposed system, it can be possible to develop a tool for Visual Analytics which will be very improved and effective. It can be possible to extend the existing system which will provide thorough clinical assessment for visual interactive knowledge discovery in large electronic health record databases.

## V. CONCLUSION

Integration of visualization techniques and more established methods combines fast automatic data mining algorithms with the intuitive power of the human mind, which improve the quality and speed of the data mining process. Visual data mining techniques are used to manage the vast amounts of relational and semi structured information, including database management and data warehouse systems. From the above discussion and paper reviewed, it can be possible to present a novel temporal event matrix representation and learning framework. It can be possible to extend the existing system with the development of an interactive tool for visualization. The ultimate goal is to bring the power of visualization technology to every desktop to allow accurate, better, faster and intuitive exploration of very large data resources.

 

**REFERENCES**


[1] Fei Wang, Noah Lee, Jianying Hu, Jimeng Sun, Shahram Ebadollahi, And Andrew F. Laine,"A Framework For Mining Signatures From Event Sequences And Its Applications In Healthcare Data", *IEEE Transactions On Pattern Analysis And Machine Intelligence*, Vol. 35, No. 2, February 2013.

[2] 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,* pp. 2689-2694, 2007.

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

[4] 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.

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

[6] J. Eggert and E. Korner, "Sparse Coding and NMF," *Proc. IEEE Int'l Joint Conf. Neural Networks*, vol. 2, pp. 2529-2533, 2004.

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

[8] C. Fevotte and J. Idier, *Algorithms for Nonnegative Matrix Factorization with the Beta-Divergence*, *arXiv*:1010. 1763, 2010.

[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] 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*, pp. 3305-3310, 2010.

[12] 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.

[13] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms," *Proc. Eighth ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery*, pp. 2-11, 2003.

[14] J. Mairal, F. Bach Inria Willow Project-Team, and G. Sapiro, "Online Learning for Matrix Factorization and Sparse Coding," *J. Machine Learning Research*, vol. 11, pp. 19-60, 2010.

[15] F. Moerchen, "Time Series Knowledge Mining Fabian," PhD thesis, 2006.

[16] F. Moerchen and D. Fradkin, "Robust Mining of Time Intervals with Semi-Interval Partial Order Patterns," *Proc. SIAM Conf. Data Mining*, pp. 315-326, 2010.

[17] F. Mo¨rchen and A. Ultsch, "Efficient Mining of Understandable Patterns from Multivariate Interval Time Series,"*Data Mining and Knowledge Discovery*, vol. 15, no. 2, pp. 181- 215, 2007.

[18] P. OGrady and B. Pearlmutter, "Discovering Convolutive Speech Phones Using Sparseness and Non-Negativity," *Proc. Seventh Int'l Conf. Independent Component Analysis and Signal Separation*, pp. 520- 527, 2007.

[19] R. Andrew Russell, "Mobile Robot Learning by Self-Observation," *Autonomous Robots*, vol. 16, no. 1, pp. 81-93, (2004).

[20] J. Shlens, G.D. Field, J.L. Gauthier, M. Greschner, A. Sher, A.M. Litke, and E.J. Chichilnisky, "The Structure of Large-Scale Synchronized Firing in Primate Retina," *J. Neuroscience: The Official J. Soc. for Neuroscience*, vol. 29, no. 15, pp. 5022-5031, 2009.

[21] P. Smaragdis, "Non-Negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs," *Proc. Fifth Int'l Conf. Independent Component Analysis and Blind Signal Separation*, 2004.

[22] Daniel A. Keim, "Information Visualization and Visual Data Mining", *IEEE Transactions on Visualization And Computer Graphics*, Vol. 7, No. 1, January-March 2002.

[23] Daniel A. Keim, Florian Mansmann, Jorn Schneidewind, and Hartmut Ziegler, "Challenges in visual data analysis", *In Proceedings of the conference on Infor-mation Visualization*, IV '06, pages 9–16, Washington, DC, USA, 2006. IEEE Computer Society.


 