ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

Mining Signatures from Event Sequences

S.H.Rajput[1], Chetan Jadhav[2], Yogesh Deshmukh[3], Sandip Sonawane[4], Hemant Jadhav[5]

Assistant Professor, Information Technology Department, SSBT COET, Jalgaon, India[1]

Student, BE.Information Technology, SSBT COET, Jalgaon, India[2,3,4,5]

Abstract: This paper proposes a novel secular knowledge representation and learning framework to proposed large-scale secular signature mining of longitudinal heterogeneous occasional data. The framework allows the presentation, extra4ction, and mining of high order latent occasion event structure and relationships between single and many sequences. The prescribed data representation maps the heterogeneous sequences to a image by encoding occasions as a structured spatial-secular shape process. We have suggested clinical assessment for naked interactive knowledge discovery in large electronic health record databases.

Keywords: Secular signature mining, sparse coding, dictionary positive matrix factorization.

I. INTRODUCTION

Data mining can be explained as an process that extracts some new nontrivial information contained in large databases. the aim is to discover hidden patterns, unexpected trends or other relationships in the data using a combination of techniques from machine study, statistics and database technologies. This new discipline today get application in a wide and range of business, scientific and engineering processes. for example, huge databases of loan applications are exist which record different kinds of personal and economical information about the applicants. These databases can be mined for difficult patterns leading to defaults which can help determine whether a future loan module must be accept or reject. several bytes of remote-sensing picture data are collected from satellites around the globe. Data mining can help release potential locations of some natural resources or assist in building fast warning systems for physical surrounding disasters like oil slick etc. other situations where data mining can be of use include analysis of medical data of hospitals in a town to conjecture, for example, potential outbreaks of infectious diseases.

II. LITERATURE SURVEY

Temporal Data Mining

Data mining is actually an integral part of Knowledge Discovery within Database (KDD) process, which is the overall process of converting malicious data into useful information [1]. A typical KDD process which consists of five steps [2]:

1) Data collection and cleaning: choosing attributes, dealing with errors, identification of the necessary background knowledge data, etc.
2) Choice of pattern discovery type: deciding on the types of knowledge to be implemented, parameter selection, etc.
3) Discovery of patterns: execute algorithms for implementing different types of patterns
4) Pattern Presentation: choosing interesting patterns, visualization of results, etc.
5) Get knowledge into use.

Temporal Data Mining (TDM) deals with the problem of mining patterns from temporal data, which should be either symbolic sequences or numerical time series. It has the capability to view for interesting correlations or rules in large sets of temporal data, which might be overviewed when the temporal component is ignored or treated as a simple numeric, attribute [3].Currently TDM is a fast expanding field with many research results reported and many new temporal data mining analysis strategy or prototypes developed recently. There are two factors that distribute to the popularity of temporal data mining. The starting factor is increase in the volume of temporal data stored, as there are too many real-world applications deal with huge amount of temporal data. The second factor is the implementing recognition in the value of temporal data. In various application domains, temporal data are now being seen as invaluable assets from which hidden knowledge can be obtained, so as to help understand the past or plan for the future [4].

TDM covers ranging spectrum of paradigms for knowledge modelling and discovery. Since temporal data mining is relatively a new field of research, there is no widely gained taxonomy yet. Several approaches have been used to categorized data mining problems and algorithms. Roddick & Spiliopoulou (2002) [3] have given a comprehensive overview of techniques for the mining of temporal data using three dimensions: data type, mining operations and type of timing information

All Temporal Sequential Pattern in Data Mining Tasks

Data mining has been used in a wide range of applications. However, the possible goals of data mining, which are often called tasks of data mining, can be grouped into some broad categories: conjecture, classification, clustering, search and retrieval, and pattern discovery [5]. This categorization follows the categorization of data mining tasks enlarged to temporal data mining [6].

Conjecture: Conjecture is the task of explicitly modelling variable dependencies to guess a subset of the variables

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

from others. The task of time series conjecture is to forecast future values of the time series based on its past samples. In order to perform the conjecture, one needs to build a predictive model from the data.

The conjecture problem for symbolic sequences has been addressed in AI research by Dietterich and Michalski (1985) [8].

Classification: Classification is the task of alloting class labels to the data according to a model learned from the training data where the classes are known. Classification has not gained much attention in temporal data mining [9]. In sequence classification, each sequence given to the system is assumed to belong to one of predefined classes and the goal is to automatically resolved the corresponding category for a given input sequence. Examples of sequence classification applications include signature verification [10], gesture recognition [11], and hand-written word recognition [12].

Clustering: Clustering is the process of finding natural groups, called clusters, in the data. Clustering of time series is concerned with grouping a collection of time series (or sequences) based on their similarity. Time series clustering has been shown effective in providing useful information in various domains [13]. For example, in financial data, clustering can be used to group stocks that exhibit same trends in price movements.Clustering of sequences is relatively less explored but is becoming increasingly important in data mining applications such as web usage mining and bioinformatics [5]. A survey on clustering time series has been presented by Liao (2005) [13].

Searching and Retrieval: Searching and retrieval are concerned with efficiently locating subsequences or sub-series in large databases of sequences or time series. In data mining, query based searches are more concerned with the problem of efficiently locating approximate matching than exact matching, known as content-based retrieval. An example of a time series retrieval application is to find out all the days of the year in which a particular stock had similar movements to those of today. Another example is finding products with similar demand cycles.
An example of sequence retrieval is finding gene expression patterns that are similar to the expression pattern of a given gene. In order to address the time series retrieval problem, different notions of similarity between time series and indexing techniques have been proposed. There is considerably less work in the area of sequence retrieval, and the problem is more general and difficult. For more detail about time series and sequence retrieval can be found in Das and Gunopulos (2003) [15].

Pattern Discovery: Unlike in search and retrieval applications, in the pattern discovery there is no specific query in hand with which to search the database. The objective is simply to discover all patterns of interest. While the other tasks described earlier have their origins in other disciplines like statistics, machine learning or pattern recognition, the pattern discovery task has its origin in data mining itself. A pattern is a local structure in the data. There are many ways of defining what constitutes a pattern. There is no universal notion for interestingness of a pattern either. However, one concept that is normally used in data mining is that of frequent patterns, that is, patterns that occurs many times in the data. Much of data mining literature is worried with formulating useful pattern structures and developing efficient algorithms for searching frequent patterns. Methods for finding frequent patterns are important because they can be used for discovering useful rules, which in turn can be used to infer some interesting regularities in the data. A rule usually consists of a pair of a lefthand side proposition (the predecessor) and a right-hand side proposition (the consequent). The rule states that when the antecedent is true, then the consequent will be true as well.

Therefore, to apply the pattern discovery methods on time series data, the time series should be first converted into a discrete representation, for example by first forming subsequences (using a sliding window) and then clustering these subsequence's using a suitable measure of pattern similarity [16].Another method can be used by quantizing the time series into levels and representing each level (e.g., high, medium, etc.) by a symbol [17].

III. EXISTING SYSTEM

Searching latent temporal signatures is important in many domains as they encode temporal ideas such as matter trends, episodes, cycles, and irregularity. For instance, in the medical region latent event signatures facilitate decision support for patient diagnosis, prognosis, and management. Particular interest is the temporal side of information hidden in event data that may be used to perform intelligent reasoning and inference about the latent relationships between event entities over time. An event institutions can be a person, an object, or a place in time. For example, in the medical domain a patient would be considered as an event institutions, where visits to the doctor's office would be considered as events.

DISADVANTAGES OF EXISTING SYSTEM:

Temporal event signature mining for knowledge discovery is a difficult problem. In this case, several problems need to be addressed:

1) The EKR(Event Knowledge Representation) should handle the time-invariant representation of multiple event institutions as two event institutions can be considered similar if they contain the same temporal signatures at different time intervals or locations,
2) EKR should be flexible to grouply represent different types of event structure such as single multivariate events and event intervals to let a rich representation of complex event relationships,
3) EKR should be scalable to support analysis and inference on big-scale databases, and
4) EKR should be dispersed to enable interpretability of the learned signatures by humans.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

d. PROPOSED SYSTEM

All This paper proposes a novel Temporal Event Matrix Representation (TEMR) and learning framework to perform temporal signature mining for big-scale longitudinal and heterogeneous event data. Basically, our TEMR framework gives the event data as a spatial-temporal matrix, where one dimension of the matrix communicates to the type of the events and the other dimension represents the 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 instinctive framework for encoding the temporal knowledge data contained in the event sequences. To improve the scalability of the forward approach, we developed an online updating technology. At last, the effectiveness of the proposed algorithm is validated on a real-world healthcare dataset.

ADVANTAGES OF PROPOSED SYSTEM:

- First, on the knowledge descriptions level, TEMR provides a visual matrix-based representation of confusing event data composed of different types of events as well as event intervals, which helps the joint representation of both continuous and discrete valued data.
- Second, on the algorithmic level, we propose a doubly dispersed convolutional matrix approximation-based formulation for searching the latent signatures within the datasets. Moreover, we obtain a multiplicative updates method to solve the problem and proved theoretically its convergence. We further propose a different academic optimization scheme for big-scale longitudinal event signature mining of multiple event entities in a group.
- Third, on the begining level, we have confimed our approach using both constructed data and a real-world Electronic Health Records (EHRs) dataset which contains the longitudinal medical records of over 20k patients over one year time. We placed the results on the detected signatures, assemble behavior of the algorithm, and the final matrix rebuilded errors.

IV. RELATED WORK

This section reviews some earlier work related to this paper, which is grouped into two parts.

The first part reviews work on the topic of knowledge representations for seculars data mining.

The second part shapes related work on nonnegative matrix factorization (NMF) and its various extensions.

2.1 Secural Knowledge Representations

There are two types of secural data, continuous and discrete. For knowledge representation of continuous time data, one of the most popular approaches is to transfer the multivariate continuous time series into discrete symbolic representations. For instance, Line al. [12] summarized existing period series representations as data adaptive, such as Piecewise Linear Approximation (PLA), Flexible Piecewise Constant Approximation (FPCA), the Singular Value Decomposition (SVD), and Symbolic Aggregate Estimation (SAE), and non-data flexible, such as the standard Discrete Fourier transform (DFT), Discrete Wavelet Transform (DWT), and Piecewise Aggregate Approximation (PAA).

For information representation of discrete time series data, Moerchen et al. [14], [16], [15] proposed a novel Time Series Knowledge Representation (TSKR) as a pattern language (grammar) for temporal knowledge discovery from multivariate time series and symbolic interval data, where the secular knowledge representation is 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 orders.

The TEMR framework we propose in this paper gives another separate way to represent the temporal knowledges contained in to the existing symbolic and grammar-based representations, our approach is more instinctive and easy to perform. Because we can always gives a matrix as an image.

V. TEMPORAL EVENT SIGNATURE MINING

Suppose we have a event matrix X 2 IRn_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 across the period axis. We call the one-side convolutional operator ?

This operator is specially composed of all events; thus there is no convolution on the vertical axis. Fig. 1 gives us an intuitive graphical illustration of the procedure of oneside convolution, where the bottom image is obtained on topleft and the time vector on top-right.

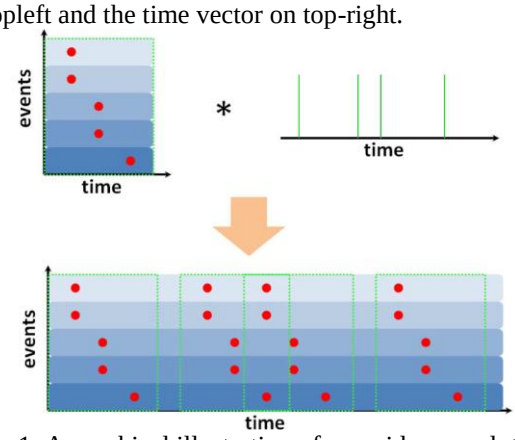

Fig. 1. A graphical illustration of one-side convolution. The top left figure shows the secular signatures, and the top right figure is the time axis, where we use green bars to diplays the position where the pattern appears.

Mining Signatures from a Single Event Sequence

Now coming back to our problem, we have the TEMR representation of the event matrix; the aim is to find the latent secural signatures from this event matrix using matrix approximation techniques.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

Mining Signatures from Multiple Event Sequences
In many real-world scenarios, we are not only interested in establishing the signatures within a single event sequence, but also in searching signatures from multiple event sequences.

In this case, it creates more sense to search signatures from a group of patients with similar disease conditions rather than a single patient. More formally, we assume the case where the event matrices are confident of n event sequences.

VI. EXPERIMENTS ON CONSTRUCTED DATA

In this section, we will present the experiments of our proposed algorithm on several constructed datasets.

Data Sets
We have created four sets of constructed datasets. Each set contains data matrices encoded with our proposed TEMR framework. All constructed data matrices have 30 rows and 120 columns. The data matrices encode events as binary activation units in the form of a single 1 or 0 valued pixel, where a value of 1 (black) denoted an event recognition and 0 (white) no event activity. Each row of the matrix refers to a particular event-type-level category and each column to a single time unit scale (e.g., days).

The first set of data is constructed to test the effectiveness of our individual OSC-NMF approach, which made of Moerchens's TSKR event-interval-test-pattern [14], [15], [16]. that has been abstracted from a tutorial figure. The pattern contained a trivariate interval event sequence, where Tones (e.g., A, B, C) represent different event interval times, Chords represent conjunction of Tones, and Phrases represent a partial ordering of the Chords. The pattern is shown in Fig. 3. The red box corresponds to the halfway ordered Phrase (AB-ABC-AC) and the green box to (AB-BC-AC) accordingly.

The right figure in Fig. 4 shows a dataset consisted of various secular concepts and operators including
1. synchronicity (red box),
2. trend of decreasing conjunction (green box),
3. trend of increasing conjunction (blue box),
4. concurrency (orange box).

Middle: Moerchen's event interval test pattern showing an alternation between Chord configurations 1 (red box) and 2 (green box). Right: From left to right, the red box shows a constructed test pattern of synchronicity, the green box shows an event pattern trend of decreasing conjunction activity, the blue box shows an event pattern trend of increasing conjunction activity, and the last orange box shows the event pattern of concurrency. Other temporal operators and concepts such as order, time, and periodicity are implicitly represented in the red, green, blue and orange enclosed patterns. Note that the constructed datasets do not have a labeled event category specified
The second dataset is developed for testing the robustness of separate OSC-NMF in the scenario where there are noisy events and different pattern elasticities contained in the data. The third set of data is created for testing the effectiveness of our group OSC-NMF method, which is consisted of three data matrices that are shown in Fig.5 The group dataset consists same and individual secular event types. The red box shows secular event pattern 1,which occurs in all three data samples. The green box shows secular event pattern 2, which also occurs in all three data samples with multiple occurrences. The blue box shows, secular event pattern 3 that only occurs in the left and middle data samples. All patterns span a duration window of seven days and an event-pattern dimension of 30.

The fourth dataset is made for examining the robustness of group OSC-NMF in the cases where there are noisy events and varying pattern flexibilty.We constructed 1,000 samples for each category, and the two types randomly appear 10 times each for every sample. For the datasets of testing noise compassions, we randomly add different levels of events to each data matrix. For the datasets of testing pattern elasticity tolerance, we randomly add 0.3 percent noisy events, and then randomly change the levels of pattern elasticity's

Testing Phase:
Testing phase involves the conjecture of unknown data sample. In checking we check those data that does not come under the dataset we have considered. After the conjecture, we will get the class labels.

Naive Bayes:
The Bayesian Classification shows a supervised learning technique as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can remove diagnostic and predictive issues. Naive Bayes algorithm is based on Bayesian Theorem.

Bayesian Theorem:
Given training data X, posterior probability of a hypothesis H, P(H|X), follows the Bayes theorem P(H|X)=P(X|H)P(H)/P(X) (1.1)

Algorithm:
The Naive Bayes algorithm is based on Bayesian theorem as given by equation (1.1) Steps in algorithm are as follows:

1) Each data sample is represented by an n dimensional feature vector, X = (x1, x2….. xn), depicting n measurements made on the sample from n attributes, respectively A1, A2, An.
2) Suppose that there are m classes, C1, C2……Cm. Given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditioned if and only if:

P(Ci/X)>P(Cj/X) for all 1< = j< = m and j!= i

Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes theorem,

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

3) As P(X) is constant for all classes, only P(X|Ci)P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, i.e. P(C1) = P(C2) = …..= P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be estimated by P(Ci) = si/s , where Si is the number of training samples of class Ci, and s is the total number of training samples. on X. That is, the naive probability assigns an unknown sample X to the class Ci [2].

VII. CONCLUSION

In this paper, we have presented a novel secular event matrix representation and learning framework in conjunction with an in-depth validation on both constructed and real world datasets. The framework has wide applicability to a variety of data and application domains that involve largescale 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 overcomplete 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.

REFERENCES

[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, pp. 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 Conjecture: Concepts, Models, and Algorithms," Math. Problems in Eng., vol. 2010, pp. 1-23, 2010.
[5]. J. Eggert and E. Korner, "Sparse Coding and NMF," Proc. IEEE Int'l Joint Conf. Neural Networks, vol. 2, pp. 2529-2533, 2004.
[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]. C. Fe´votte and J. Idier, Algorithms for Nonnegative Matrix Factorization with the Beta-Divergence, arXiv:1010.1763, 2010.
[8]. P.O. Hoyer, "Non-Negative Matrix Factorization with Sparseness Constraints," J. Machine Learning Research, vol. 5, pp. 1457-1469,2004. [9] P.O. Hoyer, "Non-Negative Sparse Coding," Proc. 12th IEEE Workshop Neural Networks for Signal Processing, 2002.
[9]. Y.R. Ramesh Kumar and P.A. Govardhan, "Stock Market Conjectures—Integrating User Perception for Extracting Better Conjecture a Framework," Int'l J. Eng. Science, vol. 2, no. 7,pp. 3305-3310, 2010.
[10]. 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.
[11]. 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.
[12]. 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.
[13]. F. Moerchen, "Time Series Knowledge Mining Fabian," PhD thesis, 2006.
[14]. 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.
[15]. 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..
[16]. 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.
[17]. R. Andrew Russell, "Mobile Robot Learning by Self-Observation,"Autonomous Robots, vol. 16, no. 1, pp. 81-93, (2004).
[18]. 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
[19]. 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.
[20]. L. Xie, H. Sundaram, and M. Campbell, "Event Mining in Multimedia Streams," Proc. IEEE, vol. 96, no. 4, pp. 623-647, Apr. 2008.
[21]. B.A. Young, E. Lin, M. Von Korff, G. Simon, P. Ciechanowski, E.J. Ludman, S. Everson-Stewart, L. Kinder, M. Oliver, E.J. Boyko, and W.J. Katon, "Diabetes Complications Severity Index and Risk of Mortality, Hospitalization, and Healthcare utilization," The Am. J. Managed Care, vol. 14, no. 1, pp. 15-23, 2008.