Temporal Data mining in healthcare: A Survey

Dinesh Kumar Bhawnani[1], Dr. Sunita Soni[2], Dr. Arpana Rawal[3]

[1]Assistant Professor, Computer Science & Engineering, BIT, Durg
[2]Professor, Computer Applications, BIT, Durg
[3]Professor, Computer Science & Engineering, BIT, Durg

Abstract: Discovering patterns in databases area fundamental data mining task, preferably the algorithms are found in abundance for discovering temporal information patterns. One of the data mining tasks on temporal database is sequential pattern mining which consists of generating useful subsequences in a set of sequences. Sequential pattern mining too has resolved many analytical issues in real life applications including bioinformatics, market basket analysis, education field, and healthcare prognosis patterns. This paper described very recent studies on sequential pattern mining applied in healthcare systems.

1. Introduction: Data mining is used to extract useful information from data stored in databases to take some decisions based on the patterns or rules generated. The major tasks in fundamental data mining are clustering, classification, regression, outlier analysis, trend analysis and pattern mining. Pattern mining is a process of inventing remarkable, useful, and unforeseen patterns in databases.

Temporal data can be of three types(Mitsa, 2010) (i) Time series, which represent ordered real valued measurements at regular temporal intervals. (ii) Temporal sequences, which can be time stamped at regular or irregular time intervals. (iii) Semantic temporal data, which are defined within the context of ontology.

Sequential pattern mining(Kumar, Krishna, Limited, & Raju, 2012)is the process of generating frequent subsequences in a sequence database. A sequence s is called as a frequent sequence or a sequential pattern if and only if its support is greater than or equal to minimum support defined by user. The support measure of a sequence s in a sequence database is defined as the number of distinct sequences which contain s. A sequence database contains a list of sequences with sequence identifiers. The major objective of sequential pattern mining is to determine useful subsequences in a sequence database, which is also called constructive sequential relationships between items.

In healthcare, the chances of doubt that the patients are really suffering from chronic diseases in patients are increasing day by day; also identifying accurate lines of treatment or reason for having chronic disease is still a challenging task.

2. Related Work:

(Ramirez, Cook, Peterson, & Peterson, 2000) introduced TEMPADIS, i.e. Temporal Pattern Discovery System, which employs Event Set Sequence method to find patterns in temporal data. They applied exploratory analysis on a customized version of the GSP Algorithm model. However the basic algorithm is the same, the details of implementation are different. The GSP algorithm was planned to develop sequences of events that either occurred or did not, where the occurrence was important to the patterns revealed. TEMPADIS uses a weakest-link/average-link approach for formulating whether or not a sequence being discussed is supported by a specified patient's data. Method was applied to the database of HIV patients and the results are presented.

(Li et al., 2005) studied an anti-monotone property for mining most favorable risk pattern sets and proposed association rule mining algorithm to make use of the property in risk pattern discovery. The method has been applied to a real world data set to find patterns related with an allergic event for ACE inhibitors. They had made utilize of an epidemiological metric, relative risk, in finding usefulness of patterns and have concluded that it is an optimal rule mining dilemma to find high risk patterns. They applied the method to a real world medical and pharmaceutical linked data set and have discovered some patterns potentially helpful in clinical practice.

(Raj, O'Connor, & Das, 2007) had presented an ontology-driven temporal mining method, called ChronoMiner. While most mining algorithms require data to be input in a single table, ChronoMiner can search for attractive temporal patterns among many input tables and at different levels of hierarchical demonstration. They had presented the relevance of method to the finding of temporal associations between recently arising mutations in the HIV genome and past drugs. They discussed various mechanism of ChronoMiner, including its user interface, and provide results of a study demonstrating the efficiency and possible value of ChronoMiner on an existing HIV drug resistance data warehouse.

(Asha, Natarajan, & Murthy, 2011) applied associative classification techniques which include CBA and CMAR for predicting tuberculosis in a databases which mainly includes twelve preliminary symptoms and one class attribute. They predict the class label of unidentified sample as Pulmonary Tuberculosis (PTB) or Retroviral Pulmonary Tuberculosis (RPTB), i.e., TB along with HIV based on higher confidence rule. Their Results showed that most of the classifier rules help in the best prediction of tuberculosis which assists doctors in their diagnosis decisions.

(Dagliati et al., 2014) presented the results of a workflow mining method to examine complex temporal datasets of Type 2 Diabetes patients.The main idea behind their approach was to use a group of temporal data mining methods in order to obtain healthcare decisions. Theirapproach includes processing raw data, derived from diverse data sources, and makes event logs, which contain significant healthcare activities. They showed how diverse data can be processed to develop a steady depiction and how due to the discovery of behavioral patterns, it is likely to rebuildsignificant clinical pathways and evaluate the complications that might arise during the process of care. The presented framework allows the classification of attractive clusters of patients with related care histories and consequently, the reassessment of their risk profiles.

(Chin et al., 2015) presented a new framework for early RA estimation that utilize data preprocessing, risk pattern mining, validation, and analysis in that two risk patterns can be revealed namely Type I which refers to well known risk patterns that have been identified by existing studies and Type II which depicts unknown association risk patterns that have seldom or never been reported in the literature. The framework had four phases; data preprocessing, risk pattern mining, risk pattern validation, and analysis. In the preprocessing phase, the data extracted is analyzed. These data include patient IDs, outpatient dates, and diagnosis codes and is divided into RA disease groups and a non-RA disease group. In the data mining and validation phases, theyutilized a ten-fold validation method. Associative classification mining is then applied to find out general RA disease risk patterns with which to generate the risk model. Theyassess the capability of the RA disease risk models by assessing their sensitivity and specificity.

(Cheng, Lin, Chiang, & Tseng, 2016)applied SPADE sequential rule mining to a General Practice database to discover rules concerning a patients age, gender and medical history. By incorporating these rules into current healthcare a patient can be highlighted as vulnerable to a future sickness based on past or current sicknesses, gender and year of birth. After applying these rules they made if then statements to predict sickness in future.

(CHENG, Lin, Chiang, & Tseng, 2017) proposed a new framework for early estimation on chronic diseases by generating sequential risk patterns with time interval information from diagnostic clinical records using combination of sequential rules mining and classification modeling methods. The proposed framework consists of four phases namely data preprocessing, risk pattern mining, classification modeling and post analysis. The diagnostic records were cleaned by removing noises during the phase of data pre-processing, including COPD patient with its diagnostic criteria and format cleaning and then imported to the sequential patternbased classification algorithm for sequential risk pattern mining and classification model building. In the post analysis phase, they filter the mined sequential risk rules by querying PubMed to assess the novelties of each rule. Furthermore, the time gap between probable symptoms and definite diagnosis of COPD is gathered from exposed risk pattern. In this way, the proposed framework enables the detection of risk patterns that are considerablyconnected with COPD and will help physicians to analyze potential COPD patients in early phases.

3. Sequence mining tool and framework:

The collection of open source implementations of sequential pattern mining algorithms is the SPMF (Sequential Pattern Mining Framework). The hyperlink to this library of tools is http://www.philippe-fournier-viger.com/spmf/. It includesapproximately one-hundred and twenty algorithms for generatingsequential patterns, sequential rules, periodic patterns, item sets and association rules in temporal databases. Thislibrary is developed in Java and is platform independent and its code is released under the GPL v3 license. It can be run as standalonesoftware or being integrated in other Java software programs.

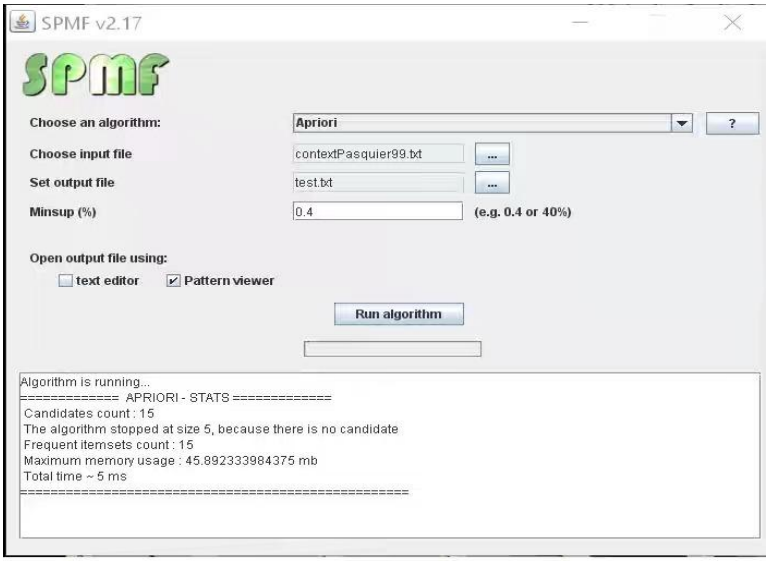

Figure 1. GUI interface of SPMF (Sequential Pattern Mining Framework)

Till date, SPMF algorithms including GSP (Ramirez et al., 2000)and SPADE (Cheng et al., 2016)have been predominantly used in many healthcare studies to discover sequential patterns in recent years. These studies have motivated us to try more SPMF algorithmic candidates for healthcare datasets. Moreover, algorithmic efficiency and practicability of these algorithms are yet to be evaluated.

4. Conclusion:

In this paper, we discussed the temporal and sequential pattern mining algorithms used in the healthcare for accurate diagnosis and prognosis of chronic diseases. There is a lot of scope still left for the enhanced scalability, high efficiency and low complexities for the algorithms due to high volume data being generated in continuous pace. The authors foresee to formulate a hybrid approach bymixingtwo or moresoft computing algorithms for early prognosis prediction of diseasesusing healthcare datasets. In summary, the field of temporal data mining in healthcare is comparatively meagerly explored and requires many new developments in the near future.

References:

Asha, T., Natarajan, S., & Murthy, K. N. B. (2011). Associative classification in the prediction of tuberculosis. *Proceedings of the International Conference & Workshop on Emerging Trends in Technology - ICWET '11*, (Icwet), 1327. https://doi.org/10.1145/1980022.1980315

CHENG, Y., Lin, Y., Chiang, K., & Tseng, V. (2017). Mining Sequential Risk Patterns from Large-Scale Clinical Databases for Early Assessment of Chronic Diseases: A Case Study on Chronic Obstructive Pulmonary Disease. *IEEE Journal of Biomedical and Health Informatics*, *2194*(c), 1–1.

https://doi.org/10.1109/JBHI.2017.2657802

Cheng, Y. T., Lin, Y. F., Chiang, K. H., & Tseng, V. S. (2016). Mining disease sequential risk patterns from nationwide clinical databases for early assessment of chronic obstructive pulmonary disease. In *3rd IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2016* (pp. 324–327). https://doi.org/10.1109/BHI.2016.7455900

Chin, C. Y., Meng, Y. W., Lin, T. C., Cheng, S. Y., Yang, Y. H. K., & Tseng, V. S. (2015). Mining disease risk patterns from nationwide clinical databases for the assessment of early rheumatoid arthritis risk. *PLoS ONE*, *10*(4), 1–20. https://doi.org/10.1371/journal.pone.0122508

Dagliati, A., Sacchi, L., Cerra, C., Leporati, P., De Cata, P., Chiovato, L., … Bellazzi, R. (2014). Temporal data mining and process mining techniques to identify cardiovascular risk-associated clinical pathways in Type 2 diabetes patients. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)* (pp. 240–243). https://doi.org/10.1109/BHI.2014.6864348

Kumar, P., Krishna, P. R., Limited, I., & Raju, S. B. (2012). *Pattern discovery using sequence data mining: applications and studies*. https://doi.org/10.4018/978-1-61350-056-9.ch011

Li, J., Fu, A. W., He, H., Chen, J., Jin, H., McAullay, D., … Kelman, C. (2005). Mining risk patterns in medical data. In *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD '05* (p. 770). https://doi.org/10.1145/1081870.1081971

Mitsa, T. (2010). *Temporal Data Mining*. *Clinics in Laboratory Medicine* (Vol. 28). https://doi.org/10.2165/11537630-000000000-00000

Raj, R., O'Connor, M. J., & Das, A. K. (2007). An ontology-driven method for hierarchical mining of temporal patterns: application to HIV drug resistance research. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, 614–9. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2655843&tool=pmcentrez&rendertype=abstract

Ramirez, J. C. G., Cook, D. J., Peterson, L. L., & Peterson, D. M. (2000). Temporal pattern discovery in course-of-disease data. *IEEE Transactions on Engineering in Medicine and Biology Magazine*, *19*(4), 63–71. https://doi.org/10.1109/51.853483