Big Data Analytics in Healthcare – Pattern Mining of Temporal Clinical Events

Svetla Boytcheva[1], Galia Angelova[1], Dimitar Tcharaktchiev[2], Zhivko Angelov[3]

[1] Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria
[2] Medical University Sofia, University Specialised Hospital for Active Treatment of Endocrinology, Bulgaria
[3] Adiss Lab Ltd., Sofia, Bulgaria

Abstract

Emails: svetla.boytcheva@gmail.com, galia@lml.bas.bg, dimitardt@gmail.com, angelov@adiss-bg.com

Keywords: Medical Informatics, Big Data, Text mining, Temporal events, Data mining

1 Motivation

Analysing relations between temporal events in clinical narratives has high importance for proving different hypothesis in healthcare: in risk factors analysis, treatment effect assessment, comparative analysis of treatment with different medications and dosage; monitoring of disease complications as well as in epidemiology for identifying complex relations between different disorders and causes for their coexistence – so called comorbidity. A lot of research efforts were reported in the area of electronic health records (EHR) visualisation and analysis of periodical data for single patient or searching patterns for a cohort of patients [4, 5, 13, 14, 15]. The work [16] proposes a method for temporal event matrix representation and a learning framework that discovers complex latent event patterns or diabetes mellitus complications. Patnaik et al [4, 5] report one of the first attempts for mining patients' history in big data scope – processing over 1.6 million of patient histories. They demonstrate a system called EMRView for mining the precedence relationships to identify and visualize partially order information. Three tasks are addressed in their research: mining parallel episodes, tracking serial extensions, and learning partial orders.

Mining frequent event pattern is a major task in data mining. It filters events with similar importance and features; this relationship can be specified by temporal constraints. There are two major tasks in data mining related to the temporal events analysis: *(i)* frequent patterns mining and *(ii)* frequent sequence mining. The difference between them is that in the first case, the event order does not matter.

In *frequent patterns mining* the events are considered as sets – collections of objects called itemsets. We investigate how often two or more objects co-occur. Usually

they are considered as a database of transactions presented like tuples $(transaction, itemset)$; the sets of transaction identifiers are called tidsets. Several methods are proposed for solving this task that vary from the naive BruteForce and Apriori algorithms, where the search space is organised as a prefix tree, to Eclat Algorithm that uses tidsets directly for support computation, by processing prefix equivalence classes [12]. An improvement of Eclat is dEclat, by reducing the space by keeping only the differences in the tidsets as opposed to the full tidsets. Another efficient algorithm is Frequent Pattern Tree Approach – FPGrowth Algorithm. Using the generated frequent patterns by all these methods we can later generate association rules.

For *frequent sequence mining* the order does matter [12]. The Level-wise generalised sequential pattern (GSP) mining algorithm searches the prefix tree using breadth-first search. SPADE algorithm applies vertical sequence mining, by recording for each symbol the position at which it occurs. PrefixSpan algorithm uses projection-based sequence mining by storing only the suffix after the first occurrence of each symbol and removing infrequent symbols from the suffix. This algorithm uses depth-first search only for the individual symbols in the projection database.

There are different mining approaches for temporal events, for instance we can consider sequences leading to certain target event [6]. Gyet and Quiniou propose recursive depth-first algorithm QTIPrefixSpan that explores the extensions of temporal patterns. Further they extract temporal sequences with quantitative temporal intervals with different models using a hyper-cube presentation and develop a version of EM algorithm for candidates' generation [8]. Patnaik et al. present the streaming algorithm for mining frequent episodes over a window of recent events in the stream [5]. Monroe et al. [11] presents a system with visual tools that allows the user to narrow iteratively the process for mining patterns to the desired target with application in EHRs. Yang et al. [10] describe another application of temporal event sequence mining for mining patient histories. They develop a model-based method for discovering common progression stages in general event sequences.

2 Project Setup

The main goal of our research is to examine comorbidity of diseases and their relationship/causality with different treatment, i.e. how the treatment of a disease can affect the co-existing other disorders. This is a quite challenging task, because the number of diagnoses (more than 10,000) and of medications (approx. 6,500) is huge. Thus the theoretically possible variations of diagnoses and corresponding treatments are above 10^{500} for one patient. That is why we shall examine separately chronic vs. acute diseases [9] and afterwards shall combine the patterns into more complex ones. Chronic diseases constitute a major cause of mortality according to the World Health Organisation (WHO) reports and their study is of higher importance for healthcare.

In order to solve this challenging task we split it down into several subtasks:

- *Mining of itemset's frequent patterns* where itemsets contain distinct chronic diseases only. Afterwards for each frequent pattern of chronic diseases we find

frequent patterns of treatment and explore their relationship in order to identify complex patterns. As a result we need to generate association rules.

- *Frequent sequences mining* – we search for causality and risk factors for chronic diseases. In this task several experiments are made with no limitations of the distance between events, only the order matters, and exploring different window limitations between events – 1 month, 3 month etc. In this task we consider also more complex sequences like parallel episodes/disorders.
- *Mining of periodical events* – searching for periodical sequences of certain acute disease in patients' histories.

3 Materials

We deal with a repository of pseudonymous Outpatient Records (OR) in Bulgarian language provided by the Bulgarian National Health Insurance Fund (NHIF) in XML format. The majority of data necessary for the health management are structured in fields with XML tags, but there are still some free-text fields that contain important explanations about the patient: "Anamnesis", "Status", "Clinical examinations" and "Therapy".

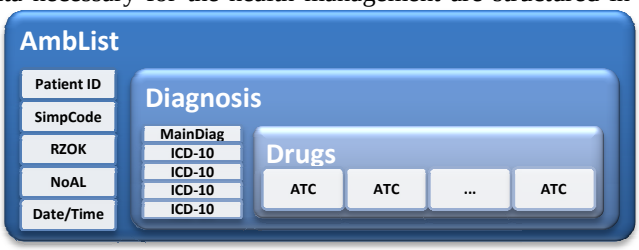

Fig. 1. Structured Event Data

From the XML fields with corresponding tags we know the Patient ID, the code of doctors' medical specialty (SimpCode), region of practice (RZOK), Date/Time and ID of the outpatient record (NoAL). XML tags also point to the main diagnose and additional diagnoses with their codes according to the International Classification of Diseases, 10th Revision (ICD-10) [3]. For extraction of information about the treatment we use a Text mining tool because the ORs contain free texts discussing drugs, dosage, frequency and route mainly in the "Therapy" section. Sometimes the "Anamnesis" also contains sentences that discuss the current or previous treatment. We developed a drug extractor using regular expressions to describe linguistic patterns [2]. There are more than 80 different patterns for matching text units to ATC drug names/codes [1] and NHIF drug codes, medication name, dosage and frequency. Currently, the extractor is elaborated and handles 2,239 drug names included in the NHIF nomenclatures.

Our experiments for pattern search are made on three collections of outpatient records that are used as training and test corpora, see Table 1. They contain data about patients suffering from Schizophrenia (ICD-10 code F20), Hyperprolactinaemia (ICD-10 code E22.1), and Diabetes Mellitus (ICD-10 codes E10-E15). These collections are of primary interest for our project because they contain cases with high diversity of chronic disorders.

Table 1. Characteristics of Data Collections

Characteristics \ Collections	S1	S2	S3
Outpatient Records	1,682,429	288,977	6,327,503
Patients	45,945	4,033	435,953
Diagnose ICD-10	F20	E22.1	E10-E15
Period	3 years	3 years	1 year
Size (GB)	4	1	18

4 Methods

Each collection $S_j,\ j = 1,2,3$ is processed independently from the other two collections. For the collection S_j the set of all different patients $P_j = \{p_{1j}, p_{2j}, \dots, p_{Nj}\}$ is extracted. For each patient $p_{ij} \in P_j$ events sequence of tuples $(event, timestamp)$ is generated: $E(p_{ij}) = \langle (e_{1j}, t_{1j}), (e_{2j}, t_{2j}), \dots, (e_{k_ij}, t_{k_ij}) \rangle,\ i = \overline{1,N}$ where $t_{n-1j} \leq t_{nj}$, $n = \overline{2,N}$.

To investigate both cases: with no limitations and with different window limitations of the distance between events, we store two versions of the temporal event sequences database for the collections $S_j, j = 1,2,3$. In the first version all timestamps are substituted with consecutive numbers starting from 0. In this case the particular dates of events do not matter, only the order matters. In the second version all time stamps are replaced with relative time – to the first event in the sequence we assign time 0, and for all other events the timestamp is converted to the number of days distance from the first event. In this case the distance between events does matter. In addition for each of these databases two subversions are generated– for chronic diseases only, and for acute diseases only. Searching patterns requires mappings between sequences illustrated in Fig. 2.

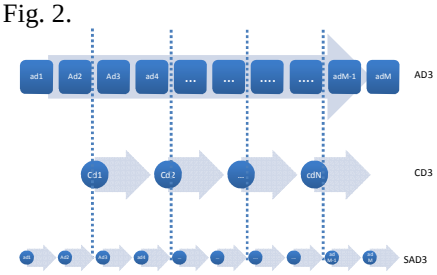

Figure 2 Mapping of the acute diseases sequence over the chronic diseases sequence

We designed a system for exploring complex relationships between disorders and treatments (See Fig. 3). It contains two main modules - for text mining and for data mining, and two repositories – for XML documents (ORs) and for structured data (temporal events sequences). The text mining module is responsible for the conversion of the raw text data concerning treatment and status [2] to structured event data and in addition for the "translation" of the structured data in the XML document to

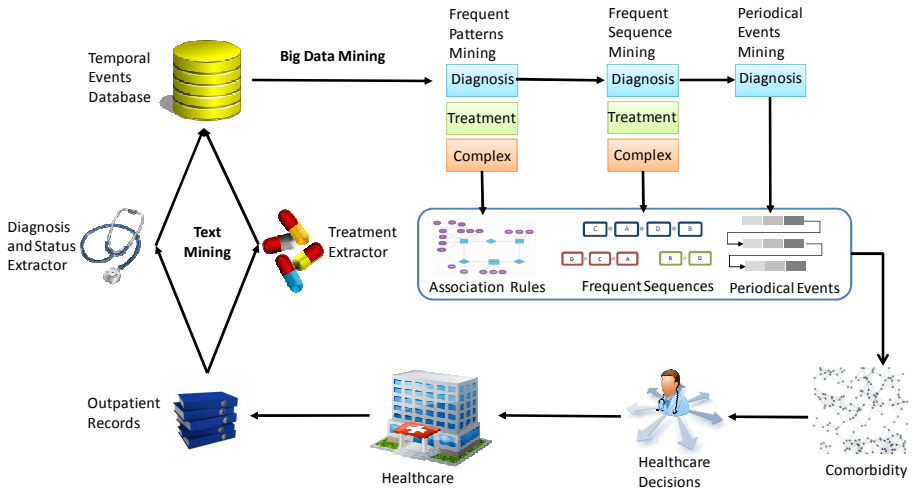

Fig. 3. System Architecture

event data. The data mining module uses a cascade approach for solving the three main tasks listed in Section 2. Initially it applies a modification of the classical Apriori algorithm [12] and generates a chronic diseases itemsets lattice and a prefix-based search tree, by also keeping partial orders for each two items and their support. The resulted lattice and partial orders information from the previous phase are used in the next phase - for frequent sequence mining. And finally, periodical events mining is applied for acute disease sequence.

5 Applications

Our system for event mining, presented briefly above, is applied for searching associations of Schizophrenia (SCZ), Diabetes Mellitus Type 2 (T2D) and Hyperprolactinemia. It is well known that patients with SCZ are at an increased risk of T2D, therefore a better understanding of the factors contributing to T2D is needed. SCZ is often treated with antipsychotic agents but the use of antipsychotics has been associated with Hyperprolactinemia, or elevated prolactin levels (a serious hormonal abnormality). Thus, given the large repository of OAs, that covers more than 5 mln citizens of Bulgaria, it is interesting to study associations and dependencies among SCZ, T2D and Hyperprolactinemia in the context of the treatment prescribed to the patients.

Regarding the treatment it is well known that the classical antipsychotics, blocking D2 dopamine receptors, lead to extrapyramidal effects related to antagonism in the nigrostriatal pathway, and Hyperprolactinaemia due to antagonism in the tuberoinfundibular pathway. In the early 1990s a new class of antipsychotics was introduced in the clinical practice with the alleged advantage of causing no or minimal extrapyramidal side effects, and the resulting potential to increase treatment

adherence. However, there are data, that some of these antipsychotics can induce diabetes, Hyperlipidaemia and weight gain.

Our study considers the presence of:

- Hyperprolactinemia in the patients with Schizophrenia,
- T2D and Schizophrenia in the patients with Hyperprolactinemia,
- T2D and Hyperprolactinemia in the patients with Schizophrenia and
- T2D in the patients with Schizophrenia and Hyperprolactinemia.

These co-morbidity facts together with the administrated medications were interpreted as temporal events and the event sequences were processed by the mining tool for pattern search. The data are extracted from more than 8 mln ORs (Table 1).

We found an increased rate of Hyperprolactinemia and T2D in patients with Schizophrenia, compared to presence of these diseases in patients without Shizophrenia. The finding is explicated in relation of the administrated treatment.

6 References

1. Anatomical Therapeutic Chemical (ATC) Classification System, http://atc.thedrugsinfo.com/
2. Boytcheva, S. Shallow Medication Extraction from Hospital Patient Records. In: Koutkias, V., J. Nies, S. Jensen, N. Maglaveras, and R. Beuscart (Eds.), *Studies in Health Technology and Informatics*, Vol. 166, IOS Press, pp. 119-128.
3. International Classification of Diseases and Related Health Problems 10th Revision. http://apps.who.int/classifications/icd10/browse/2015/en
4. Patnaik, D., L. Parida, P. Butler, B. J. Keller, N. Ramakrishnan, D. A. Hanauer. Experiences with Mining Temporal Event Sequences from Electronic Medical Records: Initial Successes and Some Challenges. In *Proc. 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'11)*, San Diego, Aug 2011, pp. 360-368.
5. Patnaik, D., S. Laxman, B. Chandramouli and N. Ramakrishnan. Efficient Episode Mining of Dynamic Event Streams, in *Proc. of the IEEE Int. Conf. on Data Mining (ICDM'12)*, Brussels, Belgium, Dec 2012, pp. 605-614.
6. Sun, X., M. Orlowska and X. Zhou, Finding Event-Oriented Patterns in Long Temporal Sequences, in *Proc. 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2003)*, Seoul, Korea, April 2003, Springer LNCS 2637, pp. 15-26.
7. Guyet, T. and R. Quiniou. Mining temporal patterns with quantitative intervals. In Zighed D., Z. Ras, and S. Tsumoto (Editors): *Proc. of the 4th Int. Workshop on Mining Complex Data, IEEE ICDM Workshop*, 2008, 10 pages.
8. Guyet, T. and R. Quiniou. Extracting temporal patterns from interval-based sequences. In *Proc. 22nd Int. Joint Conference on Artificial Intelligence*, 2011, pp.1306-1311.
9. Chronic diseases, WHO, http://www.who.int/topics/chronic_diseases/en/
10. Jaewon Yang, J. McAuley, J. Leskovec, P. LePendu, and N. Shah. Finding progression stages in time-evolving event sequences. In *Proc. of the 23rd international conference on World wide web* (WWW '14). ACM, New York, NY, USA, 2014, 783-794.
11. Monroe, M.; Rongjian Lan; Hanseung Lee; Plaisant, C., B. Shneiderman. Temporal Event Sequence Simplification, in *IEEE Transactions on Visualisation and Computer Graphics*, , 19(12), Dec. 2013, pp. 2227-2236.
12. Zaki, M. J. and Meira Wagner Jr. *Data Mining and Analysis: Fundamental Concepts and Algorithms.* Cambridge University Press, 2014.

13. Taowei David Wang, C. Plaisant, A. J. Quinn, R. Stanchak, S. Murphy, and B. Shneiderman. Aligning temporal data by sentinel events: discovering patterns in electronic health records. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '08), ACM, New York, NY, USA, 2008, 457-466.
14. Gotz, D., Fei Wang, and A. Perer. A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *Journal of Biomedical Informatics*, Vol. 48, April 2014, pp. 148-159
15. Rind, A., Taowei David Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant and B. Shneiderman, Interactive Information Visualization to Explore and Query Electronic Health Records, *Journal of Foundations and Trends® in Human–Computer Interaction* 5(3), 2013, pp 207-298.
16. Lee, N., A.F. Laine, Jianying Hu, Fei Wang, Jimeng Sun, and S. Ebadollahi. Mining electronic medical records to explore the linkage between healthcare resource utilization and disease severity in diabetic patients. *Proc. First IEEE Int. Conf. on Healthcare Informatics, Imaging and Systems Biology (HISB)*, 2011, pp. 250 – 257.