Semantically Enhanced Frequent Events Mining in Electronic Health Records

Svetla Boytcheva
Linguistic Modeling and Knowledge Processing Department
Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences, Bulgaria
Bulgaria, Sofia, 1040, 1 „15 November“ Str.
Phone: (+359 2) 979 53 33
svetla.boytcheva@gmail.com

Abstract

This paper proposes context based approach for frequent events mining (FEM) in Electronic Health Records (EHR). The majority of FEM methods do not take in consideration the context information of the analyzed dada. EHRs contain rich context information like demographic data, encounters, vital parameters, diagnoses, lab tests values, and prescribed therapy. Such information is crucial for proper interpretation of the complex temporal clinical events. Some applications in comorbidity identification, risk factors analysis and patients phenotyping are presented to illustrate the proposed method. Experiments were run on large collections of pseudoanonimized reimbursement requests submitted to the Bulgarian National Health Insurance Fund in 2010-2016 for more than 5 million citizens yearly. Effective explication of comorbidities and characterization of risk factors can fill knowledge gaps and assist informed clinical decision-making.

Keywords: Frequent Patterns Mining; Data Mining, Knowledge Discovery; Health Informatics.

1. Introduction

Investigation of events in healthcare requires development of complex models. One of the most explored problems is frequent events discovering in Electronic Health Records (EHR). Some approaches investigate the temporal nature of the events (Huang et al, 2013), i.e. frequent sequences mining (FSM). Other approaches consider the cumulative result of all events over the patient status. Such studies focus on frequent patterns identification (Shknevsky et al, 2017). The first type of research is focused on cause-effect relation and is suitable for applications like disease progress and treatment effect studies. Thus, some prediction methods can be defined on its bases. The frequent patterns mining (FPM) approaches take into account the collective effect of all complex factors for disease development. The FPM approaches are better for risk factors analysis, phenotyping and comorbidities identification. Temporal events relations analysis of EHR has higher importance for proving healthcare hypothesis: treatment effect assessment, disease complications monitoring, risk factors analysis. Currently such analyses are also used in epidemiology for identifying complex relations between different unrelated diseases – so called comorbidity and for research of rare diseases.

Majority of FSM and FPM approaches extract general templates only and do not take in consideration contextual information about extracted patterns. EHRs contain rich context information like demographic data (age, gender, and demographic region), encounters (clinic visits and hospitalizations), vitals (BMI, blood pressure), diagnoses, lab test data, and prescriptions. Such information is crucial for proper interpretation of the complex temporal clinical events, because some patterns can be valid only in certain context.

FPM and association rules generation are on primary interest in our study. This paper presents a context-based approach for FPM in EHRs.

The paper is structured as follows: Section 2 briefly overviews the research in the area; Section 3 describes the data collections of EHRs used in the study; Section 4 presents the theoretical background and formal presentation of the problem; Section 5 describes in details the proposed method for semantically enhance frequent patterns mining; Section 6 shows experimental results

and discusses the method application for searching of comorbidities and risk factors in big repository of EHRs; Section 7 contains the conclusion and sketches some plans for future work.

2. Related Work

Knowledge discovery in repositories of patient records is in focus of the clinical research (Yadav et al, 2018). Context information is organized as attributes of *itemsets* and *tidsets*. Attributes may have different organization - structured or unstructured. There are several methods of context information processing from the semantically enhanced FPM algorithms:

- Initially to extract general templates using classical FPM algorithms and then to add the context knowledge interpretation (Rabatel et al, 2013). The main disadvantage of this approach is that for large number of heterogeneous attributes with high variety of possible domain values, the number of combinations of attribute-value pairs that need to be explored exponentially grows.
- From small data collection to generate context models that are later summarized in more general models (Ziembiński, 2011).The main disadvantage of this approach is that for big collections it is hard to select representative small collection of data. Especially for collection of EHRs that is characterized by huge diversity of attribute-value pairs.
- To merge both the data (transaction itemsets) and the context as selected features and to apply data mining over the more complex data vectors (Stańczyk et al, 2017). The main disadvantage of this approach is that very long vectors need to be generated for each transaction, which is hard to be processed efficiently.

Some FPM methods are based on domain ontologies. In (Huang et al, 2013) are presented two semantics-driven FPM algorithms for EHR knowledge discovery for adverse drug effects prevention and prediction. The first algorithm is based on EHR domain ontologies and semantic data annotation with metadata. The second algorithm uses semantic hypergraph-based k-itemset generation. In (Rabatel et al, 2013) is proposed an approach in marketing domain, which takes into account not only the transactions that have been made but also various attributes associated with customers, like age, gender and etc. Attributes have a hierarchical structure and explore patterns at different levels of attributes abstraction. Rabatel et al designed algorithm Gespan and made experiments with about 100,000 product descriptions from *amazon.com*. In (Adda et al, 2005) is proposed Apriori-like FPM method enhanced with ontology and in addition with a pair of descriptive languages ─ for individuals data and for generic patterns, a generality relation between patterns. Unfortunately, such ontologies are domain and language specific and are not available for low resource languages.

Shknevsky et al address the issue of generation of contradictory symbolic time intervals patterns, caused by processing of vitals and lab test data, where many patients in the support set can have “very low” or “very high” value for some attribute. They propose an approach for time-interval relations patterns discovery using Semantic Adjacency Criterion (Shknevsky et al, 2017), which prevents the existence of potentially contradictory symbolic time intervals. This allows significant reduction (up to 97%) of the frequent patterns set that repeat with contradictory parts. Similar problem is also investigated in (Batal et al, 2013). They propose Minimal Predictive Temporal Patterns framework to generate a small set of predictive and non-spurious patterns. One of the early attempts in this direction was the algorithm PASCAL described in (Bastide et al, 2000). They propose optimization is based on pattern counting inference that relies on the concept of key patterns.

The major problem with EHRs repositories is that they can be incomplete and the data also can be noisy due to the technical errors. The timestamps of the events are uncertain, because the physicians do not know the exact occurrence time of some events. There can be a significant gap between the onset of some diseases and the first record for diagnosis in EHR made by the physician. Thus, a FPM method for dealing with temporal uncertainty was presented in (Ge et al, 2017).

Some new direction in frequent patterns mining in event sequences is *trajectories analysis*. In (Campagna & Pagh, 2010) is presented application of trajectories analyses to dataset of 2 million RFID (Radio-frequency identification) readings from baggage trolleys at Copenhagen Airport in order to identify the frequent passenger movement patterns. Influences of this idea reflect in the clinical *trajectories model*. In (Dabek & Caban, 2016) is used the following definition "A *clinical trajectory* can be defined as the path followed by patients between an initial heath state S_i such as being healthy to another state S_j such as being diagnosed with a specific clinical condition." Dabek and Caban propose a k-reversable approach for clinical trajectories modeling and present its application over a dataset of patients that have experienced mild traumatic brain injuries (Dabek & Caban, 2016). The results of experiments show that the method is effective in clustering and identifying common long-term effects associated with this injury. Jensen et al describe a methodology that allows disease trajectories of the cancer patients to be estimated from free text in EHRs (Jensen et al, 2017). The results of experiments show that about 80% of patient events can be predicted ahead in time.

Many other methods for frequent patterns discovery task solution were applied (Wang et al, 2012), like one-sided convolutional nonnegative matrix factorization, symbolic aggregate approximation, temporal abstraction approach for medical temporal patterns discovery.

Healthcare is considered as data-intensive domain and as such faces the challenges of big data processing problems. Chen and Zhang presents some directions, opportunities and challenges for big data analytics in commerce and business, society administration and scientific research (Chen & Zhang, 2014). On the other hand, medical data are quite sensitive, because they contain personal information. There are a lot of regulations and restrictions for their secondary usage for research. That is why cannot be used for cloud and distributed computing, which are considered recently as main technologies for big data analytics. Development of new scalable methods for pattern recognition in big healthcare data are required. Krumholz discusses the potential and importance of harnessing big data in healthcare for prediction, prevention and improvement of healthcare decision-making (Krumholz, 2014).

In this research, we attempt further development and combination of the ideas of data mining approaches like context based FPM and trajectories analysis. The experimental repository contains large collection of EHR, thus some modifications concerning scalability and efficiency are needed.

3. Materials

We use a data repository of about 262 million pseudonymised EHRs (outpatient records) submitted to the Bulgarian National Health Insurance Fund (NHIF) in period 2010-2016 for more than 5 million citizens yearly. The NHIF collects for reimbursement purpose all EHRs produced by General Practitioners and the Specialists from Ambulatory Care for every patient clinical visit.

The repository contains of EHRs ─ semi-structured files with predefined XML-format. Most data needed for health management are structured using standard nomenclatures, such as ICD[1] for diagnoses. Unfortunately still the majority of the important information concerning patient status and case history is provided like free text. EHRs contain paragraphs of unstructured text provided as separate XML tags (see table 1): "Anamnesis", "Status", "Clinical tests", "Prescribed treatment".

Table 1. Fields with free text in EHRs that supply input data to text mining components

#	XML field	Content
1	Anamnesis	Case history, previous treatments. Family history, risk factors
2	Status	Patient state, height, weight, BMI, blood pressure etc.
3	Clinical tests	Values of clinical examinations and lab data listed in arbitrary order
4	Prescribed treatment	Codes of drugs reimbursed by NHIF, free text descriptions of other drugs

[1] International Classification of Diseases and Related Health Problems 10th Revision. http://apps.who.int/classifications/icd10/browse/2015/en

The structured information contains date and time of the visit; pseudonymised personal data, pseudonymised visit-related information, demographic data (age, gender, and demographic region). All diagnoses are presented by ICD-10 codes and the name according the nomenclature. Each HER can contain a main diagnose and up to four additional diagnosis. The structured data can contain also a code if the patient needs special monitoring; a code concerning the need for hospitalization; several codes for planned consultations. In case the prescribed treatment is reimbursed by NHIF the medication information is also presented in structured format by NHIF drug codes, otherwise the recommended treatment description is presented as free text.

We are using raw data provided by NHIF, without of any preprocessing due to the lack of resources and annotated corpora. The text style for unstructured information is telegraphic. Usually with no punctuation and a lot of noise (some words are concatenated; there are many typos, syntax errors, etc.).

For information extraction from free text we are using text mining tools for medications (Boytcheva, 2011), vitals (BMI, blood pressure), and lab test data (Boytcheva et al, 2015).

From each EHRs is generated a patient event (see fig. 1) that contain structured information both from XML tags and extracted by Natural Language Processing (NLP) tools from free text. Each category of structured information in the patient event contains a set of attribute-value pairs.

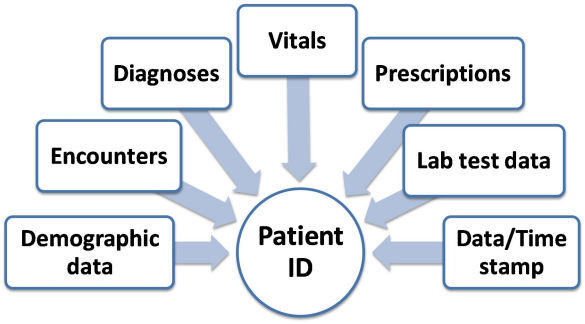

Figure 1. Patient event generated in structured form from OR of a patient single clinic visit

4. Theoretical Framework

In the classical FPM task is defined for transaction database (Aggarwal & Han, 2014) for customer transaction analysis in e-commerce. For purposes to emphasize the terminology in healthcare we will define the formal representation of the task for FPM for database of patient events.

Definition 4.1. Lets consider each patient clinic visits as a single event. The extracted set of all different patient identifiers $P = \{p_1, p_2, \dots, p_N\}$ from the collection S of EHRs is called *pids* (patient identifiers).

Definition 4.2. Let $\mathcal{E}$ be the set of all possible patient events and $\mathcal{T}$ be the set of all possible timestamps. For each patient $p_i \in P$ an event sequence of tuples $\langle event, timestamp \rangle$ is called *patient history*: $E(p_i) = \left(\langle e_1, t_1 \rangle, \langle e_2, t_2 \rangle, \dots, \langle e_{k_i}, t_{k_i} \rangle\right),\ i = \overline{1,N},\ e_j \in \mathcal{E}\ and\ t_j \in \mathcal{T}\ \ j = \overline{1,k_i},$.

Definition 4.3. Let $\mathcal{C} = \{c_1, c_2, \dots, c_p\}$ be the set of all chronic diseases[2], we call them *items*. Each subset of $X \subseteq \mathcal{C}$ is called an *itemset*.

We define a projection function (1):

$$\pi: (\mathcal{E} \times \mathcal{T})^N \to \mathcal{C}^N: \pi\big(E(p_i)\big) = C(p_i) = \left(c_{1\mathrm{i}}, c_{2\mathrm{i}}, \dots, c_{m_i}\right) \qquad (1)$$

such that for each patient $p_i \in P$ the projected time sequence contains only the first occurrence (onset) of each chronic disorder recorded in $E(p_i)$.

[2] Chronic diseases, WHO, http://www.who.int/topics/chronic_diseases/en/

Definition 4.4 Let $D \subseteq P \times \mathcal{C}$ be the set of all itemsets in our collection after projection π in the format $\langle pid, itemset \rangle$. We will call D *database*.

Definition 4.5. Let D is a database and $X \subseteq \mathcal{C}$ is an itemset. We call *support* of X in D the following set (2):

$$support(X) = \{p_i | p_i \in P, \langle p_i, Y \rangle \in D \ and \ X \subseteq Y\} \quad (2)$$

Definition 4.6. Let $\mathcal{F}$ denote the set of all frequent itemsets, i.e. $\mathcal{F} = \{X | X \subseteq \mathcal{C} \ and \ \mathrm{sup}(X) \geq minsup\}$. A frequent itemset $X \in \mathcal{F}$ is called *maximal* if it has no frequent supersets. Let $\mathcal{M}$ denote the set of all maximal frequent itemsets, i.e. $\mathcal{M} = \{X | X \in \mathcal{F} \ and \nexists Y \in \mathcal{F}, such \ that \ X \subset Y\}$.

Let 2^X denote the power set (set of all subsets) of itemset X. Then each subset of $X \in \mathcal{F}$ is also frequent itemset, i.e. $\forall Y \in 2^X \ implies \ that \ Y \in \mathcal{F}$.

Definition 4.7. For each item $c \in \mathcal{C}$ we define the set called *pidset*: $p(\mathrm{c}) = \{p_i | \langle p_i, C(p_i) \rangle \in D \ and \ \mathrm{c} \in C(p_i)\}$.

We preprocess the database D by generating pidsets and transform it to vertical database D^V: $D^V = \{\langle \mathrm{c}, p(c) \rangle | c \in \mathcal{C}\}$.

Definition 4.8. An implication in the form $I \Longrightarrow J$ is called *association rule*, where $I \subset \mathcal{C}$, $J \subset \mathcal{C}$, $I \cap J = \emptyset$. I is called *antecedent* and J is called *consequent*. The *support* of a rule is the number of pids in D that contain $I \cup J$, i.e. this is the probability

$$sup(I \Longrightarrow J) = sup(I \cup J) = P(I \cup J) \quad (3)$$

Definition 4.9. If $C\%$ of patient documents in S that contain I, contain also J, then we say that the association rule $I \Longrightarrow J$ holds with *confidence* C in S, i.e., this is the conditional probability

$$conf(I \Longrightarrow J) = P(J|I) = \frac{sup(I \Longrightarrow J)}{sup(I)} \quad (4)$$

In FPM task we are looking for itemsets $X \subseteq \mathcal{C}$ with frequency $(\mathrm{sup}(X) = |\mathrm{support(X)}|)$ above given minimal support $minsup$. In Association Rules (ARs) mining task in S is to generate all ARs with confidence above the user-defined confidence $minconf$ and support above the user-defined support $minsup$. Rules that satisfy both the $minsup$ and the $minconf$ conditions are called *strong*.

Even for reasonable pairs of values of $minsup$ and $minconf$, big datasets yield huge amounts of strong ARs and some additional filtering is needed.

Definition 4.10. The ratio of the confidence of the rule and the confidence of its consequents called *lift* that is defined as:

$$lift(I \Longrightarrow J) = \frac{P(I \cup J)}{P(I)P(J)} \quad (5)$$

The lift represents the strength of the relation between the consequent and its antecedent. Lift value less than 1 indicates independence between them. Lift value greater than 1 means that the antecedent and consequent appear together more often than expected, i.e., are correlated. Such rules are potentially useful for predicting the consequent in new sets.

5. Semantically Enhanced Frequent Patterns Mining

The collection processing is performed in three phases: preprocessing, data analysis and prediction & prevention models generation (Figure 2). After the preprocessing phase is applied data analysis over the collection of patient events in structured form. The diagnoses are considered as patient event data in focus. We use the master template approach to which context information is

subsequently added. The FPM and AR generation algorithms are applied for itemsets of ICD-10 codes of diagnoses. For ARs generation, we use algorithms for mining all association rules with the lift measure in a transaction database (Agrawal & Srikant, 1994) with implementation at SPMF[3]. For experiments, we used an algorithm for All Association Rules with FPGrowth with lift (Han et al, 2004). The generated frequent patterns itemsets (FPI) represents so called comorbidities. We need to study the nature of comorbidities and the context in which they are valid.

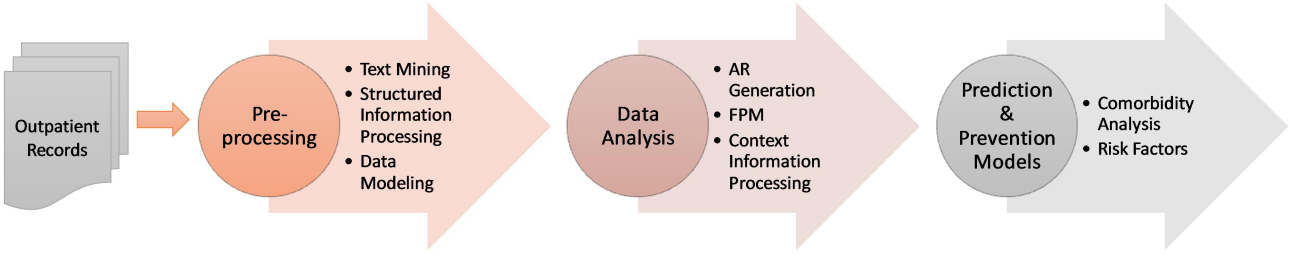

Figure 2. System Architecture

To cope with high heterogeneity of the attributes and the complexity of the hyperspace some methods like selection of focus attributes, aggregation and discretization are applied over the context data.

We define an ordered set of attributes of interest (focus attributes) $A = \{a_1, a_2, \dots, a_k\}$. The attributes are listed in decreasing order of their weight. In order to decrease the number of possible values of attributes we apply some discretization of data, i.e. using categorical values instead of numeric values. Such values in numeric ranges are mapped over categories. For instance, age value is categorized according to the World Health Organization (WHO) standard age groups. Data for body mass index (BMI) are also categorized according to the WHO[4] standard classification - *underweight*, *normal weight*, *overweight*, *obesity*. For some data concerning demographic information, like region ID we have large number of distinct values. We use aggregation on different levels, concerning background information for the region – e.g. whether it is *south*, *north*, *west*, *east*, *central*, *northwest* etc., and *mountain*, *river*, *sea*, *thermal spring*, *urban region* etc.

Definition 5.1. Context Q for some patient $p_i \in P$ is defined as the set of attribute-value pairs $\langle attribute, value \rangle$ from patient history: $E(p_i) = \left(\langle e_1, t_1 \rangle, \langle e_2, t_2 \rangle, \dots, \langle e_{k_i}, t_{k_i} \rangle\right)$, $i = \overline{1, N}$, and focus attributes set $A = \{a_1, a_2, \dots, a_k\}$. For each attribute is taken the value for the latest occurrence in the event sequence.

$$Q(p_i) = \{\langle a_1, q_1 \rangle, \langle a_2, q_2 \rangle, \dots, \langle a_k, q_k \rangle\} \quad (6)$$

We propose new cascade method *ContextFPM* for generating the context of FPI:

(1) Initially generate the set $\mathcal{M}$ of FPI for diagnosis and their support.
(2) Initialize $B \leftarrow \emptyset$
(3) Select the first attribute $a \in A$, and remove it from A.
(4) The collection S is clustered for all possible values v in the domain $D(a)$ of the attribute a. Foreach $v \in D(a)$ find $support(a, v) = \{p_i | \, p_i \in P, \langle a, v \rangle \in Q(p_i)\}$.
(5) Apply reduction $D'(a)$ of $D(a)$ for those values v of the attribute a for which $|support(a, v)| < minsup$.
(6) Foreach $\langle m, support(m) \rangle \in \mathcal{M}$

[3] http://www.philippe-fournier-viger.com/spmf/index.php?link=algorithms.php
[4] WHO, BMI Classification http://apps.who.int/bmi/index.jsp?introPage=intro_3.html

Foreach $v \in D'(a)$
find $\mathcal{F} = support(m) \cap support(a, v)$
if $|\mathcal{F}| \geq minsup$ then
if $\exists\, r = \langle m, \mathcal{F}', context\rangle \in B$
then $B \leftarrow B \cup \langle m, \mathcal{F}, context \cup \{\langle a, v\rangle\}\rangle$ *and* $\mathcal{M}' \leftarrow \langle m, \mathcal{F}\rangle$
else $B \leftarrow B \cup \langle m, \mathcal{F}, \{\langle a, v\rangle\}\rangle$ *and* $\mathcal{M}' \leftarrow \langle m, \mathcal{F}\rangle$

(7) Replace $\mathcal{M} \leftarrow \mathcal{M}'$
(8) If $A = \emptyset$ then return B and stop
else goto step (3).

6. Experiments and Results

In epidemiology, Type 2 Diabetes Mellitus (T2DM) is considered as one of the primary causes for mortality with rapidly increasing levels of prevalence each year (Zimmet et al, 2014). Thus, the study of its risk factors is with higher importance for prevention and healthcare policies improvement. In (Bellou et al, 2018) is shown the complexity of risk factors for T2DM, which include not only genetics, but lifestyle, dietary and environmental factors as well.

For experiments, we apply retrospective analysis of patients in pre-diabetes condition to identify some risk factors for T2DM. Due to the short period of the EHRs available in the repository, we cannot apply analysis over 2-year period (2013-2014) for patient with onset of the T2DM in 2015. This allows validating the generated results for risk factors.

The FPI are generated for $minsup = 0.1$. For this experiment the itemsets contain all diagnosis (both for acute and chronic diseases) from patient history. The results of data analysis for are shown in table 2.

Table 2. Collection S for patients in pre-diabetes condition in the period 2013-2014

period	2013	2014	2013-2014
EHRs	267,194	296,129	556,323
Patients	27,082	27,902	29,205
ICD-10 codes	4,701	4,834	5,503
Frequent Itemsets	7,452	8,935	32,093
Association Rules	58,299	78,052	381,012

The main comorbidity classes before the context analysis are shown in fig. 4. The highest picks in fig. 4 correspond to cardiovascular diseases. It is well known that diseases of the circulatory system are primary risk factors for T2DM: Hypertensive diseases (ICD-10 codes I10-I15), Ischaemic heart diseases (ICD-10 codes I20-I25), Atrial fibrillation and flutter (ICD-10 codes I48). Other comorbidities include Diseases of the eye and adnexa (ICD-10 codes H00-H59) and Diseases of the musculoskeletal system and connective tissue (ICD-10 codes M00-M99).

We consider as focus attributes age, sex, blood pressure, blood glucose levels and glycated hemoglobin (HbA1C) levels. The distribution of sex is 41% male and 59% female. These values have no significant deviation from the standard distribution of the population. Some age specific FPI that were identified concern different types of cancer.

The onset of T2DM is usually at age after 45. The patients with age in range 0-14 these patients were excluded after context analysis due to the support below the $minsup$ (fig. 3). For patient with age in range 15-44 the main comorbidities considered as risk factors include Obesity and Hypertension. Some correlation between the patient distributions in different age categories according to WHO can be seen in fig. 3 for patients in prediabetes condition (left) and patients with T2DM (right).

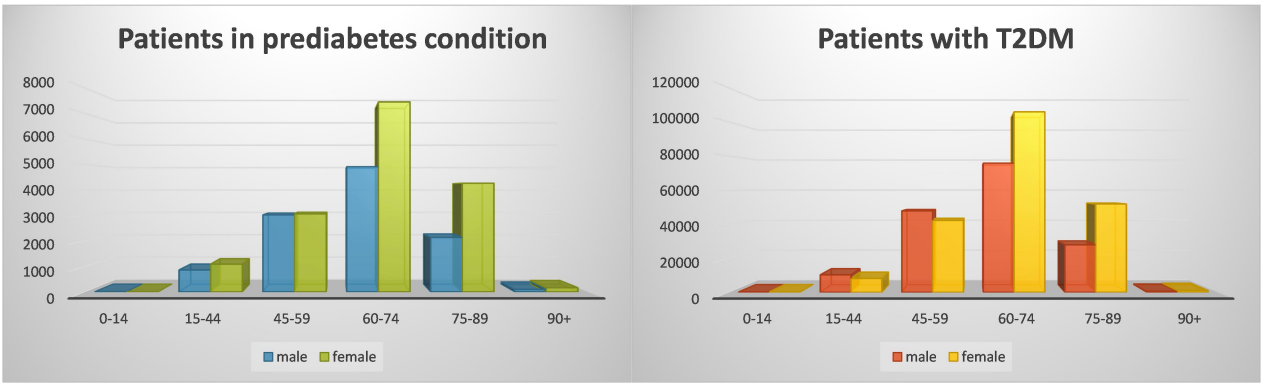

Figure 3. Age of the patients in the collection S, grouped by WHO categories (left) and the age of patients with T2DM for the population in the same period and categories (right)

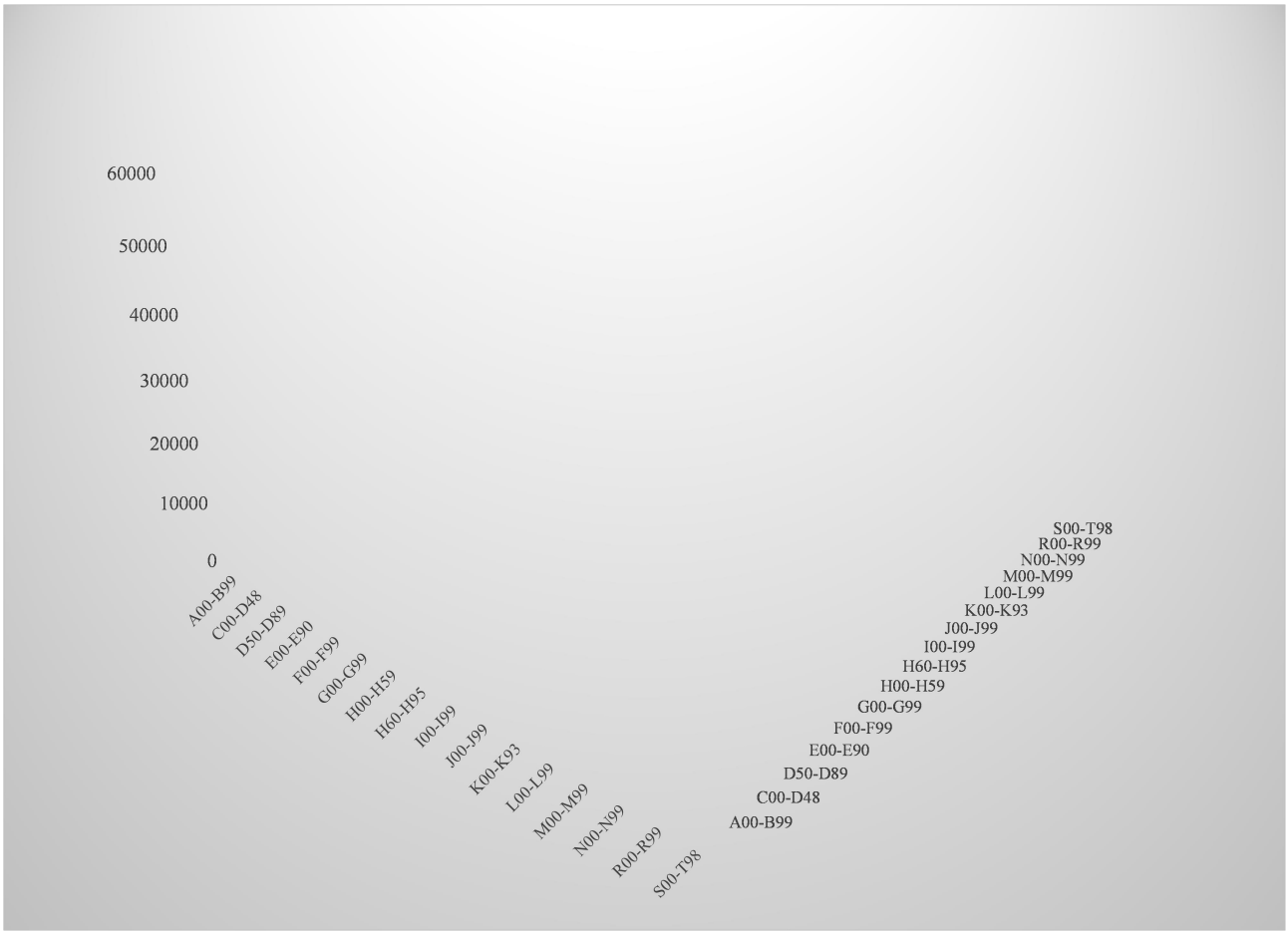

Figure 4. Comorbidities for 2013-2014 collection of EHRs grouped by classes in ICD-10

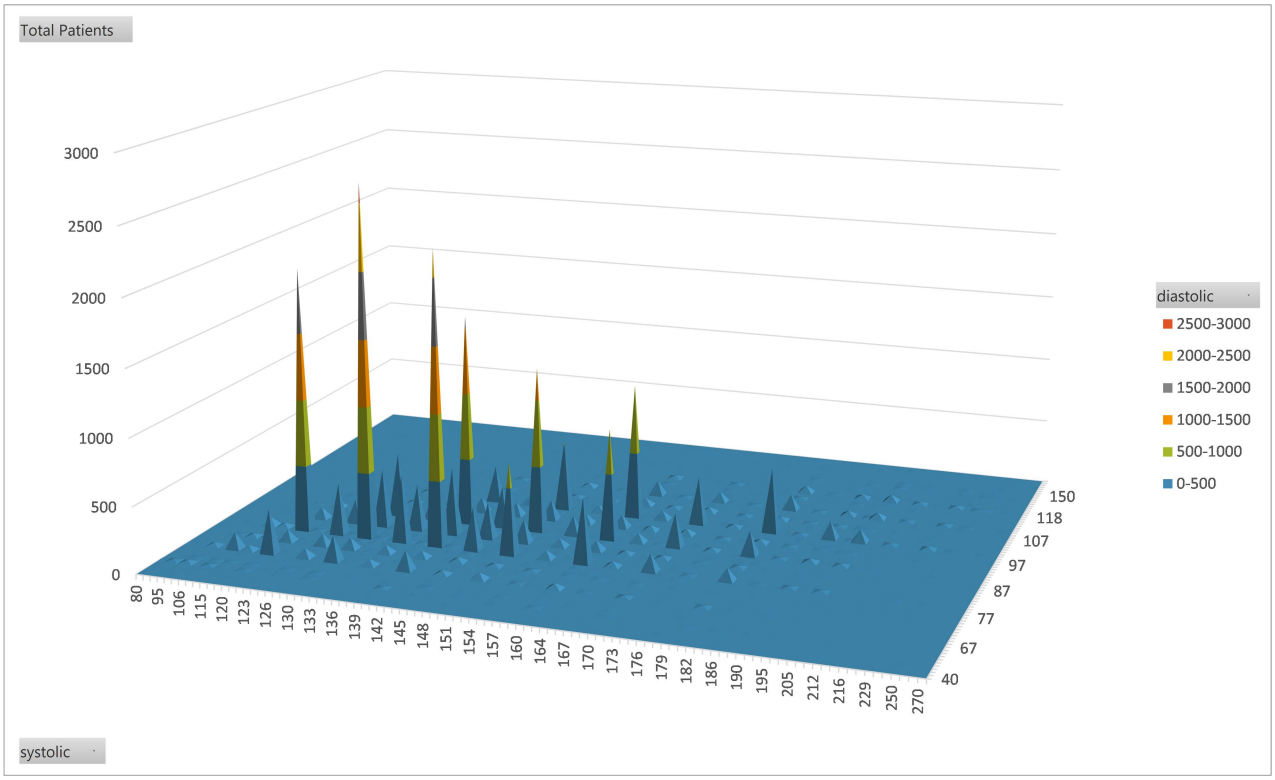

Figure 5. Blood Pressure (systolic and diastolic) levels of the patients in the collection S

For the attributes, concerning vitals we made experiments with data for blood pressure only that are available for majority of patients. The repository contains data from multiple clinic visits, but we process only data from the latest visit, that are more relevant to the event of T2DM diagnosis. For all patients data are presented for blood pressure (RR – Riva Roci) with (systolic and diastolic) levels. This allows their discretization in several categories were considered as attribute values: hypertension stage 1, hypertension stage 2, elevated blood pressure, hypotension, blood pressure in norm. The results (fig. 5) show that for more than 50% of patients have high blood pressure levels.

For the attributes concerning lab test results, we lack of data (fig. 6 and fig 7) for majority of patients, because these clinical results are usually monitored for patients with diagnosis T2DM. The tests were done for some patients at the end of 2014 immediately before the period of T2DM diagnostization. Only for patients with high levels of glucose levels and glycated hemoglobin were discovered context-based FPI. The other categories were excluded due to the low support.

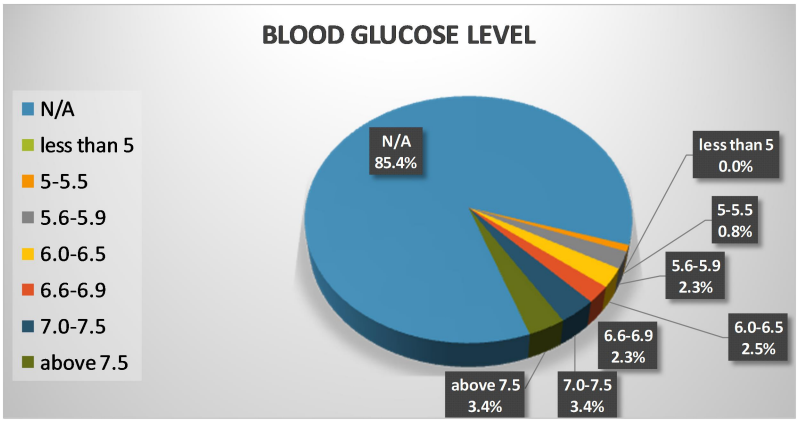

Figure 6. Blood glucose levels of the patients in the collection S, grouped by categories

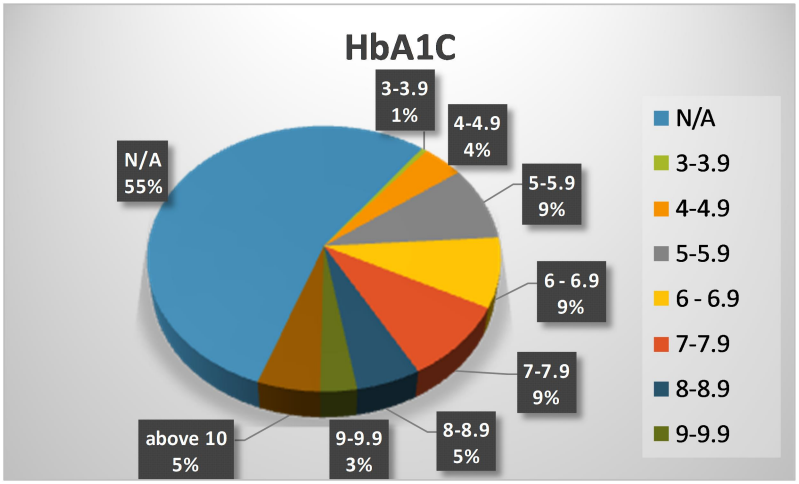

Figure 7. HbA1C levels of the patients in the collection S, grouped by categories

7. Conclusion and Further Work

In this paper, we present semantically enhanced approach *ContextFPM* for FPM of EHRs. The task is challenging, because we need apply automatic NLP in large scale for medical records for low resource language, which is novelty in this area. The secondary use of EHRs leads to knowledge discovery. New results were identified considering some types of cancer and diseases of the musculoskeletal system as risk factors for T2DM that need further investigation and explanation. We demonstrated that the context plays important role for comorbidities validation. In our previous research (Boytcheva et al, 2017) we also addressed the problem for context based FPM for EHRs. The proposed method uses support vector machines (SVM) and optimization based on block minimization method described in (Yu et al, 2012). The proposed method *ContextFPM* at is more flexible and allows partial interpretation of the context of FPI by iterative steps.

Future work includes further elaboration of specific algorithms that take into consideration temporal sequences of events. Development of more efficient tools for filtering of the generated pool of association rules. Implementation of visualization interactive tools will provide to experts functionality to explore and investigate in more details FPI findings.

Acknowledgements

The research presented here is partially supported by the grant 02/4 SpecialIZed Data MIning MethoDs Based on Semantic Attributes (IZIDA), funded by the Bulgarian National Science Fund in 2017–2019. The support of Medical University – Sofia, the Bulgarian Ministry of Health and the National Health Insurance Fund is acknowledged.

References

Adda, M., Valtchev, P., Missaoui, R., and Djeraba, C. (2005, December). On The Discovery of Semantically Enhanced Sequential Patterns. In Proceedings of the Fourth International Conference on Machine Learning and Applications, pp. 383-390. IEEE Computer Society.

Aggarwal, C. C., and Han, J. (Eds.). (2014). *Frequent pattern mining*. Springer.

Agrawal, R., and Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB, 1215, 487-499.

Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., and Lakhal, L. (2000). Mining frequent patterns with counting inference. ACM SIGKDD Explorations Newsletter, 2(2), pp. 66-75.

Batal, I., Valizadegan, H., Cooper, G. F., and Hauskrecht, M. (2013). A temporal pattern mining approach for classifying electronic health record data. ACM Transactions on Intelligent Systems and Technology (TIST), 4(4), pp. 63.

Bellou, V., Belbasis, L., Tzoulaki, I., and Evangelou, E. (2018). Risk factors for type 2 diabetes mellitus: An exposure-wide umbrella review of meta-analyses. PloS one, 13(3), e0194127.

Boytcheva, S. (2011). Shallow medication extraction from hospital patient records. Studies in health technology and informatics, 166, 119-128.

Boytcheva, S., Angelova, G., Angelov, Z., and Tcharaktchiev, D. (2015). Text mining and big data analytics for retrospective analysis of clinical texts from outpatient care. Cybernetics and Information Technologies, 15(4), 58-77.

Boytcheva, S., Angelova, G., Angelov, Z., and Tcharaktchiev, D. (2017). Mining comorbidity patterns using retrospective analysis of big collection of outpatient records. Health information science and systems, 5(1), 3.

Campagna, A., and Pagh, R. (2010, December). On finding frequent patterns in event sequences. In Proceedings of 2010 IEEE 10th International Conference on Data Mining (ICDM), pp. 755-760. IEEE.

Chen, C. P., and Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, 275, 314-347.

Dabek, F. J., and Caban, J. J. (2016). A k-reversible approach to model clinical trajectories. In AMIA Annual Symposium Proceedings, 2016, pp. 460. American Medical Informatics Association.

Ge, J., Xia, Y., Wang, J., Nadungodage, C. H., and Prabhakar, S. (2017). Sequential pattern mining in databases with temporal uncertainty. Knowledge and Information Systems, volume 51(3), pp. 821-850.

Han, J., Pei, J., Yin, Y., and Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data mining and knowledge discovery, 8(1), 53-87.

Huang, J., Huan, J., Tropsha, A., Dang, J., Zhang, H., and Xiong, M. (2013, December). Semantics-driven frequent data pattern mining on electronic health records for effective adverse drug event monitoring. In Proceedings of 2013 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 608-611. IEEE.

Jensen, K., Soguero-Ruiz, C., Mikalsen, K. O., Lindsetmo, R. O., Kouskoumvekaki, I., Girolami, M., ... and Augestad, K. M. (2017). Analysis of free text in electronic health records for identification of cancer patient trajectories. Scientific reports, 7, 46226.

Krumholz, HM. (2014). Big Data and New Knowledge in Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. Health affairs (Project Hope). 2014; 33(7), 1163-1170. doi:10.1377/hlthaff.2014.0053.

Rabatel, J., Bringay, S., and Poncelet, P. (2013). Mining sequential patterns: a context-aware approach. In Advances in Knowledge Discovery and Management, pp. 23-41. Springer, Berlin, Heidelberg.

Shknevsky, A., Shahar, Y., and Moskovitch, R. (2017). Consistent discovery of frequent interval-based temporal patterns in chronic patients' data. Journal of biomedical informatics, 75, 83-95.

Stańczyk, U., Zielosko, B., and Jain, L. C. (2017). Advances in Feature Selection for Data and Pattern Recognition, 138. Springer.

Wang, F., Lee, N., Hu, J., Sun, J., and Ebadollahi, S. (2012, August). Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 453-461. ACM.

Yadav, P., Steinbach, M., Kumar, V., and Simon, G. (2018). Mining Electronic Health Records (EHRs): A Survey. ACM Computing Surveys (CSUR), 50(6), pp. 85.

Yu, H. F., Hsieh, C. J., Chang, K. W., and Lin, C. J. (2012). Large linear classification when data cannot fit in memory. ACM Transactions on Knowledge Discovery from Data (TKDD), 5(4), 23.

Ziembiński, R. Z. (2011). Accuracy of generalized context patterns in the context based sequential patterns mining. Control and Cybernetics, 40, 585-603.
Zimmet, P. Z., Magliano, D. J., Herman, W. H., and Shaw, J. E. (2014). Diabetes: a 21st century challenge. The lancet Diabetes & endocrinology, 2(1), 56-64.

Svetla Boytcheva received her MSc in Mathematics and PhD in Computer Science from Sofia University "St. Kliment Ohridski", Bulgaria. Now she is associate professor of computer science in Linguistic Modeling and Knowledge Processing Department, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences. Her current research interests include different aspects of Artificial Intelligence and Health Informatics. She has (co-) authored 10 books and more than 70 scientific papers.