Boytcheva *et al. Health Inf Sci Syst (2017) 5:3*
https://doi.org/10.1007/s13755-017-0024-y

Health Information Science and Systems

RESEARCH

Mining comorbidity patterns using retrospective analysis of big collection of outpatient records

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

Abstract

Background: Studying comorbidities of disorders is important for detection and prevention. For discovering frequent patterns of diseases we can use retrospective analysis of population data, by filtering events with common properties and similar significance. Most frequent pattern mining methods do not consider contextual information about extracted patterns. Further data mining developments might enable more efficient applications in specific tasks like comorbidities identification.

Methods: We propose a cascade data mining approach for frequent pattern mining enriched with context information, including a new algorithm MIxCO for maximal frequent patterns mining. Text mining tools extract entities from free text and deliver additional context attributes beyond the structured information about the patients.

Results: The proposed approach was tested using pseudonymised reimbursement requests (outpatient records) submitted to the Bulgarian National Health Insurance Fund in 2010–2016 for more than 5 million citizens yearly. Experiments were run on 3 data collections. Some known comorbidities of Schizophrenia, Hyperprolactinemia and Diabetes Mellitus Type 2 are confirmed; novel hypotheses about stable comorbidities are generated. The evaluation shows that MIxCO is efficient for big dense datasets.

Conclusion: Explicating maximal frequent itemsets enables to build hypotheses concerning the relationships between the exogeneous and endogeneous factors triggering the formation of these sets. MixCO will help to identify risk groups of patients with a predisposition to develop socially-significant disorders like diabetes. This will turn static archives like the Diabetes Register in Bulgaria to a powerful alerting and predictive framework.

Keywords: Data mining, Maximal frequent patterns mining, Natural language processing, Comorbidity

Motivation

Studying comorbidities of disorders is important for detection and prevention. For discovering frequent patterns of diseases we can use retrospective analysis of population data, by filtering events with common properties and similar significance. Two major approaches to pattern search are: *(i)* frequent pattern mining (FPM) viewing the events (objects) as unordered sets and *(ii)* frequent sequence mining (FSM). Most FPM and FSM methods do not consider contextual information about extracted patterns; usually they build a prefix tree, which is huge and difficult to manipulate in memory when big data is processed [1]. Further developments of data mining (DM) might enable more efficient applications in specific tasks e.g. identification of relations between different unrelated diseases—so called comorbidity.

Clinical narratives are an underused data source that has much greater research potential than is currently realized. Biomedical scientists increasingly use text mining to extract important entities from medical texts and integrate them in various DM studies. However automated analysis of clinical texts is most successful for English while only fragmented components exist for the other languages. So advancing the biomedical Natural

*Correspondence: svetla.boytcheva@gmail.com
[1] Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria
Full list of author information is available at the end of the article

Language Processing (NLP) for under-resourced languages is another challenge to be met in order to improve DM achievements in knowledge discovery.

We propose a cascade data mining approach for frequent pattern mining enriched with context information, including a new algorithm MIxCO for maximal frequent patterns mining. Text mining tools extract entities from free text and deliver additional context attributes beyond the structured information about the patients. NLP for Bulgarian delivers entities from Outpatient Records (ORs) free texts. Novel hypotheses are generated to discover stable comorbidities and to confirm known ones. The experiments explicate some population specific comorbidities. We also discuss the effects of age, gender and demographics on these comorbidities.

The paper is structured as follows. Section 2 presents related work, Sect. 3—the data we use, Sect. 4—the methods. Section 5 discusses current experiments and their medical interpretation. Section 6 sketches further work and the conclusion.

Related work

The concept of frequent itemsets is introduced by Agrawal et al. [2]. Methods for solving FPM vary from the naive BruteForce and Apriori algorithms, where the search space is organized as a prefix tree, to Eclat/dEclat algorithm that uses tidsets directly for support computation, by processing prefix equivalence classes [1]. Another efficient algorithm is FPGrowth (Frequent Pattern Tree Approach). Using the generated frequent patterns we can later generate association rules. Most FPM algorithms generate all possible frequent patterns (FPs). The search space grows exponentially with the number of items. Summarized information for data relations can be extracted as maximal frequent itemsets (MFI). The condensed information not only accelerates the process, reducing redundancy, but also decreases significantly the number of FPs for post-analysis. All classic algorithms for FPM can be modified for MFI search, by checking for maximality at each step. There are some especially designed algorithms for MFI search, e.g. the MFCS algorithm which combines top-down and bottom-up [3]. The GenMax algorithm that uses a vertical database, diffsets and optimizations by checking whether the union of all itemsets is included already in some maximal itemset and then pruning the branch [4]. The FPMax algorithm is based on FP-trees by extending FP-growth algorithm [5]. MAFIA uses depth-first traversal of the itemset lattice with effective pruning mechanisms which is quite good especially when the database itemsets are very long [6].

NLP for English clinical texts made significant progress in algorithm development and resource construction since 2000. Open-source tools like cTAKES[1] extracts information from clinical free text. Another open source system is HITEx (Health Information Text Extraction) which extracts variables of interest from narratives [7]. Despite the limitations, the NLP importance as a supporting technology will grow due to its constant improvements [8].

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

XML field	Content
Anamnesis	Case history, previous treatments. Family history, risk factors
Status	Patient state, height, weight, BMI, blood pressure etc.
Clinical tests	Lab data and clinical examinations values in arbitrary order
Treatment	Codes of drugs reimbursed by NHIF, free texts for other drugs

Studies on multimorbidity are a great challenge given the mismatch between the high prevalence of this condition and relatively smaller number of research papers [9], which is partly due to lack of data. Machine learning (ML) is the basic technology used in such studies. For instance, four ML techniques (logistic regression, k-nearest neighbors, multifactor dimensionality reduction and support vector machines) were applied to assess risks for diabetes, hypertension and their comorbidity in a cohort of 270,172 hospital visitors (89,858 diabetic, 58,745 hypertensive and 30,522 comorbid patients) in Kuwait, with accuracy > 85% (for diabetes) and > 90% (for hypertension) [10]. An original approach for predicting a comorbid medical condition incidence and progression of medical conditions, using self-posted data available on patient-oriented social media sites, is presented in [11]. The similarity between patient postings is calculated and the risk of a condition is derived thus producing a ranked list of medical conditions for each patient. An algorithm to build medical condition progression trajectories is suggested. The condition incidence model predicts future conditions with coverage of 48% (top-20) and 75% (top-100).

Materials

The data repository we use currently contains more than 262 million pseudonymised ORs submitted to the Bulgarian National Health Insurance Fund (NHIF) in 2010-2016 for more than 5 mln citizens yearly. In Bulgaria ORs are produced by the General Practitioners and the Specialists from Ambulatory Care for every contact with the patient. Despite their primary accounting purpose the

[1] Clinical Text Analysis and Knowledge Extraction System: http://ctakes.apache.org/.

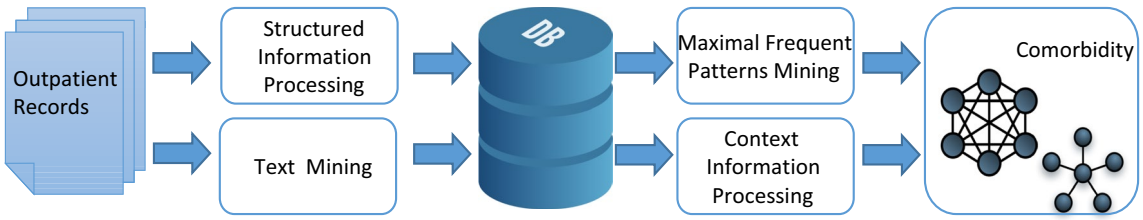

Fig. 1 System architecture

ORs summarise sufficiently the case and motivate the requested reimbursement. ORs are semi-structured files with predefined XML-format. The structured XML fields provide useful information: date and time of the visit; pseudonymised personal data, age, gender; pseudonymised visit-related information; diagnoses in ICD-10[2]; NHIF drug codes for medications that are reimbursed; a code if the patient needs special monitoring; a code concerning the need for hospitalization; several codes for planned consultations, lab tests and medical imaging. The free text OR fields, listed in Table 1, are processed by our NLP tools.

Methods

The system architecture is shown on Fig. 1. Text mining modules convert the raw text to structured data. We developed a drug extractor using regular expressions to describe linguistic patterns [12], it handles 2239 drug names included in the NHIF nomenclatures. For extraction of clinical test data (body mass index—BMI, weight, blood pressure etc.) we designed a Numerical value extractor [13].

We search for as many as possible associations between chronic diseases.[3] A tabular method using a vertical database is proposed, with depth-first traversal as well as set intersection and diffsets. Further processing of the MFI is applied to remove diagnostic related groups. Some context information is added to each MFI to study comorbidities. This information is presented as attribute-value tuples for each patient; the post-processing identifies the importance of different attributes for each MFI.

Mining maximal frequent itemsets

For the collection S of ORs we extract the set of all different patient identifiers $P = \{p_1, p_2, \ldots, p_N\}$. This set corresponds to transaction identifiers (*tids*) and we call them *pids* (patient identifiers). We consider each patient visit to a doctor as a single event. For each patient $p_i \in P$ an event sequence of tuples $\langle event, timestamp\rangle$ is generated: $E(p_i) = \left(\langle e_1, t_1\rangle, \langle e_2, t_2\rangle, \ldots, \langle e_{k_i}, t_{k_i}\rangle\right), i = \overline{1,N}$. Let $\mathcal{E}$ be the set of all possible events and $\mathcal{T}$ be the set of all possible timestamps. Let $\mathcal{C} = \{c_1, c_2, \ldots, c_p\}$ be the set of all chronic diseases, which we call *items*. Each subset of $X \subseteq \mathcal{C}$ is called an *itemset*. We define a projection function $\pi : (\mathcal{E} \times \mathcal{T})^N \rightarrow \mathcal{C}^N$: $\pi(E(p_i)) = C(p_i) = (c_{1\mathrm{i}}, c_{2\mathrm{i}}, \ldots, c_{\mathrm{mi}})$, 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 $E(p_i)$. Let $D \subseteq P \times 2^{\mathcal{C}}$ be the set of all itemsets in our collection after projection π in the format $\langle pid, itemset\rangle$. We will call D *database*. We are looking for itemsets $X \subseteq \mathcal{C}$ with frequency $(\sup(X))$ above given *minsup*. Let $\mathcal{F}$ denote the set of all frequent itemsets, i.e. $\mathcal{F} = \{X | X \subseteq \mathcal{C}$ *and* $\sup(X) \geq$ *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* $Y \in \mathcal{F}$. 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: $D^V = \langle\{\mathrm{c}, p(c)\rangle | c \in \mathcal{C}\}$. Let $w \in \mathcal{C}$, we define projection P_w of the database D^V by pidsets intersection: $P_w\left(D^V\right) = \{\langle \mathrm{c}, p'(c)\rangle | \langle \mathrm{c}, p(c)\rangle \in D^V, \mathrm{c} \neq \mathrm{w}$ *and* $p'(c) = p(c) \cap \mathrm{p(w)}\}$ and its complement by pidsets difference: $\overline{P_w\left(D^V\right)} = \{\langle \mathrm{c}, p''(c)\rangle | \langle \mathrm{c}, p(c)\rangle \in D^V, \mathrm{c} \neq \mathrm{w}$ *and* $p''(c) = p(c) - \mathrm{p(w)}\}$. Let $f(c)$ denotes the frequency of item $c \in \mathcal{C}$ in database D^V. An item $w \in \mathcal{C}$ is called *weak*, if there has no item in $\mathcal{C}$ with support lower than $f(w)$, i.e. $\nexists c \in \mathcal{C}$ such that $f(c) < f(w)$.

Algorithm MIxCO (MIning COmorbidity)

Assume that the set of all maximal frequent itemsets $\mathcal{M}$ is initially the empty set. We reduce the database D^V by

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

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

deleting all tuples that contain items with support below *minsup* and process further the obtained database $D^{V'}$. Obviously the maximal frequent itemsets will contain as many as possible items, thus they must contain also items with low frequency. In order to identify maximal frequent itemsets we start from the weakest item $w \in \mathcal{C}$ in $D^{V'}$. There are two cases: either a maximal frequent itemset X contains w, or it does not contain it. Thus we need to split $D^{V'}$ in two subsets by projections $P_w\left(D^{V'}\right)$ and $\overline{P_w\left(D^{V'}\right)}$. We apply recursively the algorithm MIxCO for searching all maximal frequent itemsets in $P_w\left(D^{V'}\right)$. Let the result set of all maximal frequent itemsets in $P_w\left(D^{V'}\right)$ be $\mathcal{M}_w$. We add to each of them the item w: $\mathcal{M}'_w = \{Y | X \in \mathcal{M}_w, Y = X \cup \{w\}\}$ and obtain the maximal frequent itemsets that contain w. Let $\mathcal{B}$ be the set of all members of $P_w\left(D^{V'}\right)$ that were reduced from the algorithm MIxCO due to low frequency (bellow the *minsup*). These items cannot be reduced from further considerations because they have low frequency in combination with w, but support above *minsup* in the entire database $D^{V'}$ and they can be members of maximal frequent itemsets that do not contain w. We update $\overline{P_w\left(D^{V'}\right)}$ by adding those itemsets that contain members of $\mathcal{B}$:

$\overline{P_w\left(D^{V'}\right)}^U = \overline{P_w\left(D^{V'}\right)} \cup \{\langle c, y \cup (z \cap p(x))\rangle | \langle x, p(x)\rangle \in \mathcal{B}, \langle c, z\rangle \in P_w\left(D^{V'}\right), \langle c, y\rangle \in \overline{P_w\left(D^{V'}\right)}\}$. We apply recursively the algorithm MIxCO for searching all maximal frequent itemsets in $\overline{P_w\left(D^{V'}\right)}^U$. Let the result set of all maximal frequent itemsets in $\overline{P_w\left(D^{V'}\right)}^U$ be $\overline{\mathcal{M}_w}$. Then the result set of all maximal frequent itemsets of the database D is the union $\mathcal{M} = \mathcal{M}'_w \cup \overline{\mathcal{M}_w}$. Finally we reduce $\mathcal{M}$ by removing all frequent patterns that are not maximal, if any.

We illustrate *MIxCO* by a synthetic example (Fig. 2). Itemsets of ICD-10 codes for 10 patients are presented. For each ICD-10 code (F20, E11, I11, M17, I20, E66, J44) is generated a set of pids, i.e. DV. We apply reduction for $minsup = 3$ and obtain $B = \{\langle M17, \{2, 4\}\rangle, \langle E66, \{10\}\rangle, \langle J44, \{5\}\rangle\}$. The weakest item of the new set DV' is $w = I20$. On the next step we partition DV' into two subsets by projection $P_{I20}\left(DV'\right)$ and $\overline{P_{I20}(DV')}$. First we start processing $P_{I20}\left(DV'\right)$ and apply reduction with $B' = \{\langle \text{I11}, \{8\}\rangle\}$. The weakest item in the resuced set $P_{I20}\left(DV''\right)$ is $w = F20$. We apply projection and obtain to subsets $P_{F20}\left(P_{I20}\left(DV''\right)\right)$ and $\overline{P_{F20}(P_{I20}(DV''))}$. Because for $P_{F20}\left(P_{I20}\left(DV''\right)\right)$ no reduction is applied and its cardinality is 1, we return the frequent itemset $\mathcal{M} = \{\{F20, E11, I20\}\}$, which contains items from both projections F20 and I20 and the only left item E11 in the later subset. The subset $\overline{P_{F20}(P_{I20}(DV''))}$ is empty and the algorithm terminates processing the subset $P_{I20}\left(DV'\right)$. We continue by processing $\overline{P_{I20}(DV')}$ and update it by the reduced data from B'. No further reductions are applied to the updated set $\overline{P_{I20}(DV'')}$, because all subsets have support above minsup. The weakest item in $\overline{P_{I20}(DV'')}$ is $w = F20$. We apply projection and obtain to subsets $P_{F20}\left(\overline{P_{I20}(DV'')}\right)$ and $\overline{P_{F20}\left(\overline{P_{I20}(DV'')}\right)}$. For $P_{F20}\left(\overline{P_{I20}(DV'')}\right)$ no reduction is applied and its weakest item is $w = E11$. We apply projection and obtain to subsets $P_{E11}\left(P_{F20}\left(\overline{P_{I20}(DV'')}\right)\right)$ and $\overline{P_{E11}\left(P_{F20}\left(\overline{P_{I20}(DV'')}\right)\right)}$ and so on. The frequent itemsets $\mathcal{M} = \{\{F20, E11, I20\}, \{F20, E11\}, \{F20, I11\}, \{E11, I11\}\}$, produced at the end, are presented in oval shapes in the leaves of the tree on Fig. 2. Finally we reduce the non maximal itemsets from $\mathcal{M}$, i.e. $\{F20, E11\} \subset \{F20, E11, I20\}$, presented in shadow in Fig. 2. The result set $\mathcal{M} = \{\{F20, E11, I20\}, \{F20, I11\}, \{E11, I11\}\}$ contain maximal frequent itemsets only.

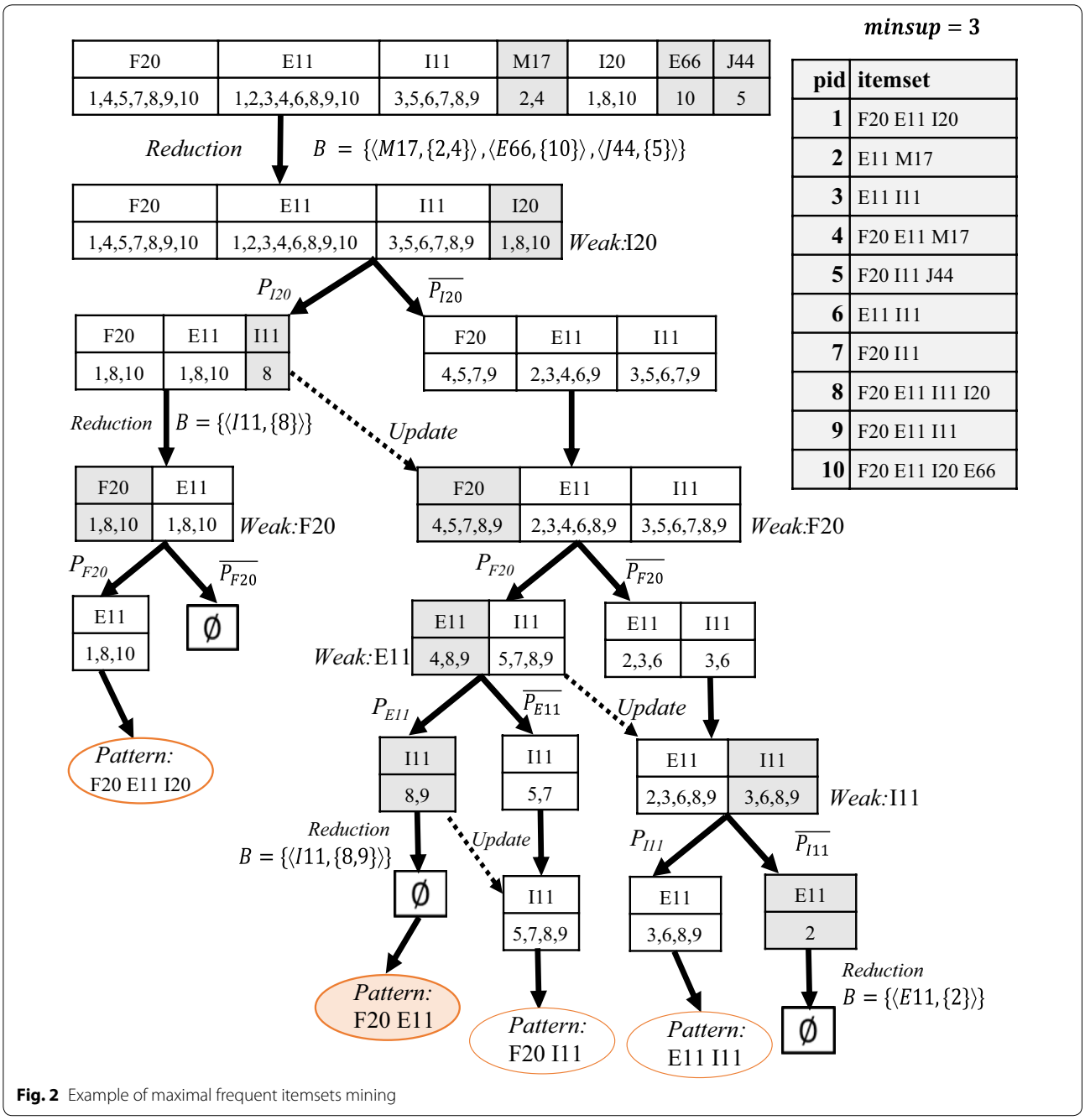

Fig. 2 Example of maximal frequent itemsets mining

Algorithm MIxCO

//Initial Call: $\mathcal{M} \leftarrow \emptyset$, $\mathcal{B} \leftarrow \emptyset$, $DV \leftarrow \{\langle c, p(c)\rangle | c \in \mathcal{C}\}$,

$MIxCO(\mathcal{M}, DV, \text{minsup}, \mathcal{B})$

1 **foreach** $\langle c, p(c)\rangle \in DV$ // reduction
2 **if** $|p(c)| < minsup$ **then** $\mathcal{B} \leftarrow \langle c, p(c)\rangle$ and $DV \leftarrow DV - \langle c, p(c)\rangle$
3 **if** $(DV == \emptyset)$ **then return**
4 **if** $(|DV = \{\langle c, p(c)\rangle\}| == 1)$ **then** $\mathcal{M} \leftarrow c$ and **return**
5 Find the weakest item w in DV
6 $DV = P_w(DV) \cup \overline{P_w(DV)}$
7 $\mathcal{M}_w \leftarrow \emptyset$
8 $\mathcal{B}_w \leftarrow \emptyset$
9 $MIxCO(\mathcal{M}_w, P_w(DV), \text{minsup}, \mathcal{B}_w)$
10 **foreach** $\langle c, y\rangle \in \overline{P_w(DV)}$ // update
11 **foreach** $\langle x, p(x)\rangle \in \mathcal{B}_w$
12 find $\langle c, z\rangle \in P_w(DV)$
13 $y \leftarrow y \cup (z \cap p(x))$
14 $\overline{\mathcal{M}_w} \leftarrow \emptyset$
15 $\overline{\mathcal{B}_w} \leftarrow \emptyset$
16 $MIxCO(\overline{\mathcal{M}_w}, \overline{P_w(DV)}, \text{minsup}, \overline{\mathcal{B}_w})$
17 $\mathcal{B} \leftarrow \mathcal{B} \cup \overline{\mathcal{B}_w}$
18 **foreach** $m \in \mathcal{M}_w$
19 $m \leftarrow m \cup \{w\}$
20 $\mathcal{M} \leftarrow \mathcal{M}_w \cup \overline{\mathcal{M}_w}$
21 Reduce non maximal itemsets from $\mathcal{M}$

Context information

Comorbidities need to be studied in the context where they occur so we add semantic attributes to each event—patient demographics, age and gender, treatment, status etc.

We define a set of attributes of interest $A = \{a_1, a_2, \ldots, a_k\}$. Context Q for some patient $p_i \in P$ is defined as the set of attribute-value pairs from patient profile information: $Q(p_i) = \{\langle a_1, q_1\rangle, \langle a_2, q_2\rangle, \ldots, \langle a_k, q_k\rangle\}$. In order to decrease the number of possible values of attributes we apply some aggregation of data. For instance age value is categorized according to the World Health Organization (WHO) standard age groups.[4] Data for body mass index (BMI) are also categorized according to the WHO[5] standard classification—*underweight*, *normal weight*, *overweight*, *obesity*. For some data concerning demographic information, like region ID we have large number of distinct values. For such data we add also some additional properties 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.

From $Q(p_i)$ we generate a feature vector $v(p_i) = (v_{1i}, v_{2i}, \ldots, v_{mi})$, where each attribute $a_j \in A$ with N_j possible values is represented by N_j consecutive positions in the vector. For the set of maximal frequent itemset $\mathcal{M}$ with cardinality $|\mathcal{M}| = K$ we have K classes of comorbidities. We apply classification of multiple classes in order to generate rules for each comorbidity class. We use SVM and optimization based on block minimization method described by Yu et al. [14].

Experiments and medical relevance

MIxCO algorithm evaluation

Some evaluation experiments were performed for MixCO and FPMax algorithms with two databases A and B. The number of transaction in both collections is 11,345, but A is very dense, and in contrast B is very sparse. The number of items in A is 4337, and in B is 3412. Table 2 shows the execution time in milliseconds for a relative minsup between 0.01 and 0.05.

[4] WHO, Standard age groups http://www.who.int/healthinfo/paper31.pdf.

[5] WHO, BMI Classification http://apps.who.int/bmi/index.jsp?introPage=intro_3.html.

Table 2 The execution time in milliseconds for experiments with two synthetic datasets

Database	Algorithm	minsup				
		0.01	0.02	0.03	0.04	0.05
A	FP-Max	9,952,681	596,652	100,899	31,148	**13,461**
	MIxCO	**45,276**	**35,700**	**32,456**	**29,026**	27,315
B	FP-Max	**112**	**94**	**87**	**81**	**77**
	MIxCO	9879	7066	5913	5255	4456

The best results in experiments are highlighted in bold

The evaluation results show that FP-Max outperforms MIxCO for big sparse databases. In contrast MIxCO shows better results for big dense databases.

Comorbidity identification

The term "comorbid" here means "indicating two or more medical conditions existing simultaneously regardless of their causal relationship". One comprehensive study of the possible relations between comorbid diseases is [15]. The authors describe 13 comorbid models, also known as "NK models", which allow to examine the etiology of the comorbidity between disorders and to predict mortality and other outcomes.

Our experiments for pattern search are made on five OR collections that are used as training and test corpora (Table 3). They contain data about patients suffering from Schizophrenia (ICD-10 code *F*20), Hyperprolactinaemia (ICD-10 code *E*22.1), and Diabetes Mellitus Type 2 (ICD-10 code *E*11). Schizophrenia and Diabetes Mellitus are chronic diseases with a variety of complications that are also chronic diseases. The collections are extracted by using a Business Intelligence Tool (BITool) [13] from the repository of ORs for approx. 5 million patients for a 3-years period.

The minsup value was set as relative minsup function of the ration between the number of patients and ORs. It is approximately between 0.015% for S2 and S3 and 0.005% for S1. This is a rather small minsup value that will guarantee coverage even for more rare chronic diseases but with sufficient support.

The noise in the data is not taken into account. We do not discuss the correctness of the clinical data from medical point of view. The average number of ORs per patient is distributed almost evenly in the collections S1–S3: 12.2 (set S1), 9.85 (S2) and 14.5 (S3) and each patient has several visits each year. On the other hand the collections are almost complete and cover the population in Bulgaria for these period.

The experimental collections were carefully selected. The association between Schizophrenia, Hyperprolactinemia, and Diabetes Mellitus Type 2 is well known so it is easier to assess the novelty of discovered comorbidities corresponding to the extracted maximal frequency itemsets.

Comorbidity interpretation in psychiatric diseases has specific aspects because in mental health comorbidity does not necessarily imply the presence of multiple diseases. It usually is the result of imprecisely distinguished mental illnesses and inability to supply a single diagnosis that accounts for all symptoms. For example in collection S1 the support of itemset $\{F20, F31\}$ is 871, where $F31$ is Bipolar affective disorder and $F20$ is Schizophrenia. Despite this imperfection, we see that the longest maximal frequent itemsets overcome this problem. Table 4 contains diseases with ICD-10 codes $I11$ (Hypertensive heart disease with heart failure), $I20$ (Angina pectoris), $I50$ (Heart failure), $I69$ (Sequelae of cerebrovascular disease). The result is not quite surprising due to the

Table 3 Characteristics of data collections

Collection	S1	S2	S3
Main diagnosis	Schizophrenia	Hyperprolactinaemia	Diabetes Mellitus Type 2
ICD-10 code	F20	E22.1	E11
Minsup(Relative)	80(0.005%)	45(0.015%)	1092 (0.017%)
Patients number	45,945	9777	435,953
ORs	1,682,429	288,977	6,327,503
Period	3 years	3 years	1 year
Total MFI	204	316	305
Longest MFI	6	5	6
ICD-10 codes	5790	4697	5473
Chronic diseases	227	228	228

Table 4 The longest maximal frequent itemsets for Schizophrenia with size = 6

Maximal frequent itemset	Support
$\{E11, F20, I11, I20, I50, I69\}$	100
$\{F20, I11, I20, I48, I50, I69\}$	83

well-studied comorbidity between Schizophrenia and Cardio-vascular diseases [15].

Interesting and unexpected results were found in the set of maximal frequent itemsets with size 5 (Table 5)—comorbidity with *M*17 Gonarthrosis (arthrosis of knee).

This correlation seems to be a new hypothesis: a search PubMed found only 3 papers referring to relations between delusions and physical diseases such as knee osteoarthritis. Even more interesting results were obtained after adding context information. The demographic data show some relation between comorbidity of $\{F20, M17\}$ and location of thermal springs in Bulgaria (Fig. 3). Expected BMI values of these patients are high but most of them have normal BMI or a little overweigh. Thus, contextualizing the FPM findings, the proposed technology supports discovery and exploration of novel correlations between phenotypes and comorbidity.

The role of phenotype for comorbidity of various diseases is known. For instance, the most often psychiatric disorder—depression—is comorbid with anxiety disorders, abuse with psychoactive substances, alcohol and drug dependence. High comorbidity is established between depression and somatic dysfunctions as well, e.g. 22–33% of the patients hospitalised for treatment of somatic diseases have depressive disorders too [16]. It is accepted that the predisposition to the development of certain disease is due to the contribution of multiple genes with little effect. The correlation between the genetic fingerprint and the environment works in both directions: people with genetic predisposition can develop certain illness when they live in the respective environment; on the other hand the genes can change the individual sensitivity to the environmental factors and contribute to the development of predisposition [17].

The experiments presented here show that deeper understanding of the interrelations between comorbidity, phenotypes and environmental factors can be achieved by finer tuning of the classical data mining techniques in order to discover unknown correlations between data items in patient records and contextual information.

Table 5 Some maximal frequent itemsets for Schizophrenia with size = 5

Maximal frequent itemset	Support
$\{E11, F20, I11, I20, M17\}$	133
$\{F20, I11, I20, I50, M17\}$	125
$\{F20, I11, I20, I69, M17\}$	108

Conclusion and future work

This paper presents a novel algorithm MixCO for MFI mining. The main advantage of MixCO is that it can process efficiently big dense datasets for small relative minsup values. This is a bottom-up approach which eliminates at the beginning the most critical items that are highly possible to be reduced in the MFI. The expected application impact of MixCO is significant. The explication of maximal frequent itemsets enables to build hypotheses concerning the causality relationships among the exogeneous and endogeneous factors that trigger the formation of these sets. Mining of patterns is shown here, and mining sequences is the next task in our agenda.

Future work includes also in-depth experiments with various OR subsets and evaluation of the effectiveness of MixCO.

The diagnoses with several possible ICD-10 codes or similar diagnoses are also not interpreted in this model. This is an important issue and we plan further investigation of it in our future work.

Finally we note that the technology can be successfully used for explication of risk groups of patients that have predisposition to develop socially-significant disorders like diabetes. This is possible given the large repository of

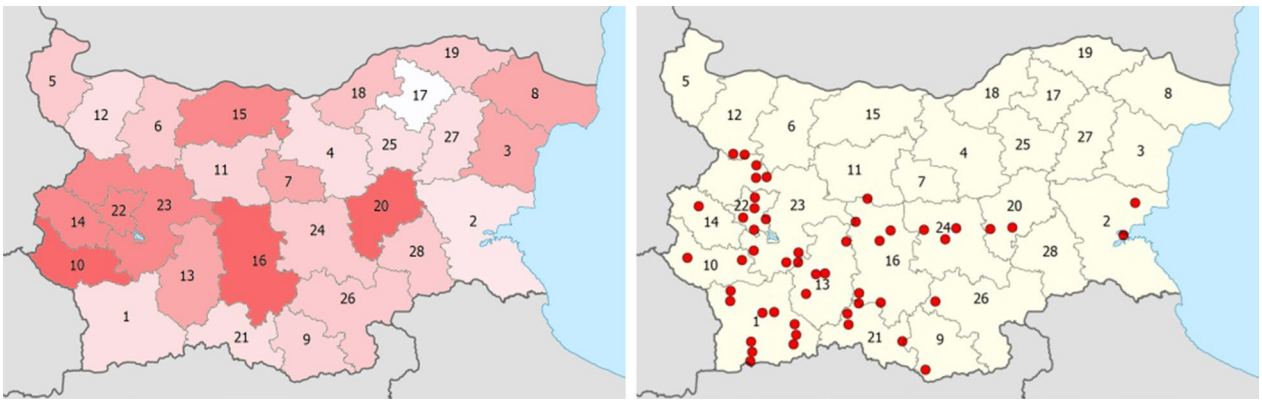

Fig. 3 Map of Bulgaria with comorbidity of $\{F20, M17\}$ (left) and Thermal springs (right)

patient-related data organised now in a national Diabetes Register for Bulgaria.[6] In this way advanced DM algorithms like MixCO and their application to repositories like the Diabetes Register in Bulgaria will turn static archives to powerful alerting and predictive frameworks.

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

Acknowledgements
The research work presented in this paper is partially supported by the grant *SpeciallZed Data MIning MethoDs Based on Semantic Attributes* (IZIDA), funded by the Bulgarian National Science Fund in 2017–2019. The team acknowledges also the support of Medical University—Sofia, the Bulgarian Ministry of Health and the Bulgarian National Health Insurance Fund.

Publisher's Note

Received: 10 August 2017 Accepted: 22 September 2017
Published online: 28 September 2017

References

1. Nasreen S, Azam MA, Shehzad K, Naeem U, Ghazanfar MA. Frequent pattern mining algorithms for finding associated frequent patterns for data streams: a survey. Procedia Comput Sci. 2014;37:109–16.
2. Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Proc. ACM Int. Conf. on management of data (SIGMOD), 1993, p. 207–216.
3. Dao-I Lin, Kedem. Zvi M.: Pincer search: a new algorithm for discovering the maximum frequent set, in advances in database technology. In: Proc. of the 6th international conference on extending database technology (EDBT'98), 1998, p. 105–119.
4. Gouda K, Zaki MJ. GenMax: an efficient algorithm for mining maximal frequent itemsets. Data Min Knowl Disc. 2005;11(3):223–42.
5. Grahne G, Zhu J. Efficiently using prefix-trees in mining frequent itemsets. In: FIMI 90, November 2003.
6. Burdick D, et al. MAFIA: a maximal frequent itemset algorithm. IEEE Trans Knowl Data Eng. 2005;17(11):1490–504.
7. Goryachev S, Sordo M, Zeng QT. A suite of natural language processing tools developed for the I2B2 project. In: AMIA annu symp proc, vol. 931, 2006.
8. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, Gainer VS, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. Br Med J. 2015;350(1):h1885.
9. Xu X, Mishra GD, Jones M. Mapping the global research landscape and knowledge gaps on multimorbidity: a bibliometric study. J Glob Health. 2017;7(1):010414.
10. Farran B, Channanath AM, Behbehani K, Thanaraj TA. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait–a cohort study. BMJ Open. 2013;3(5):e002457.
11. Ji X, Chun SA, Geller J. Predicting comorbid conditions and trajectories using social health records. IEEE Trans Nanobiosci. 2016;15(4):371–9.
12. Boytcheva S. Shallow medication extraction from hospital patient records. In: Koutkias V, Niès J, Jensen S, Maglaveras N, Beuscart R, editors. Studies in health technology and informatics series, vol. 166, IOS Press; 2011. p. 119–128.
13. Boytcheva S, Angelova G, Angelov Z, Tcharaktchiev D. Text mining and big data analytics for retrospective analysis of clinical texts from outpatient care. Cybern Inf Technol. 2015;15(4):58–77.
14. Yu H, Hsieh C, Chang K, Lin C. Large linear classification when data cannot fit in memory. ACM Trans Knowl Discov Data (TKDD). 2012;5(4):23.
15. Neale MC, Kendler KS. Models of comorbidity for multifactorial disorders. Am J Hum Genet. 1995;57:935–53.
16. Di Matteo MR, Lepper HS, Croghan TW. Depression is a risk factor for noncompliance with medical treatment: meta-analysis of the effects of anxiety and depression on patient adherence. Arch Intern Med. 2000;60(14):2101–7.
17. Milanova V. Affective disorders: interrelation between genetic and psychosocial factors. Psychosom Med. 2006;14:137–58 **(in Bulgarian language)**.

[6] http://usbale.com/Register_Diabetes.htm.