Symptoms investigation by means of Formal Concept Analysis for enhancing medical diagnoses

Christian Săcărea, Diana Şotropa, Diana Troancă

Babeş-Bolyai University of Cluj-Napoca
csacarea@math.ubbcluj.ro, diana.halita@ubbcluj.ro, dianat@cs.ubbcluj.ro

SoftCOM 2017

Overview

1. Motivation
2. Prerequisites
3. Data Collection and Analysis
4. Data Interpretation
5. Conclusions and Future Work

Knowledge discovery in medical data

Motivation:

- identifying valid, novel, useful and ultimately understandable patterns from medical data
- help the process of medical diagnosis become more accurate and efficient
- design decision support systems that can aid in achieving clinical tests at a reduced cost
- avoid clinical decision based just on doctors' intuition and experience which can lead to unwanted biases, errors and excessive medical costs

Formal Concept Analysis

Definition

Formal Context: *triple* $\mathbb{K} = (G, M, I)$ *with*

- *set* G *of* ***objects***
- *set* M *of* ***attributes***
- ***incidence relation*** $I \subseteq G \times M$

Definition

Formal Concept: *pair* (A, B) *with*

- ***extent*** $A \subseteq G$
- ***intent*** $B \subseteq M$
- $A \times B \subseteq I$
- *maximal w.r.t. this property*

Graph Databases - *Neo4j*

Graph Databases:

- data stored in the form of nodes, edges and properties
- straightforward graph representation with a direct visualization of relations among data
- as opposed to relational databases, the focus is on the nodes as well as on the relationships between the nodes

Neo4j:

- highly scalable graph database
- represents relations among the data in a flexible way
- easy to manage and to query

Medical Data being Analyzed

- records collected from a teaching hospital in Romania
- the data contains personal characteristics, symptoms and diagnostics of patients from an investigation in the department for ear, nose and throat disorders

Using Formal Concept Analysis and Neo4j on Medical Data

- we show how new knowledge about medical investigations can be discovered, by following 3 steps:
  1. finding concepts
  2. finding relations between concepts
  3. building knowledge concept lattices
- we switch the perspective and interpret the data as knowledge graphs ⇒ we infer more information from the data graphs

Formal Concept Analysis Techniques

- in order to apply FCA techniques, datasets are interpreted as many-valued contexts
- first, data is binarized using conceptual scaling
- then, data is interpreted as formal contexts: datasets with a relation between the elements
- conceptual landscapes of knowledge are built using the FCA Tools Bundle system → it makes exploration and navigation possible

Fig. 1. Correlation between principal and secondary diagnostics

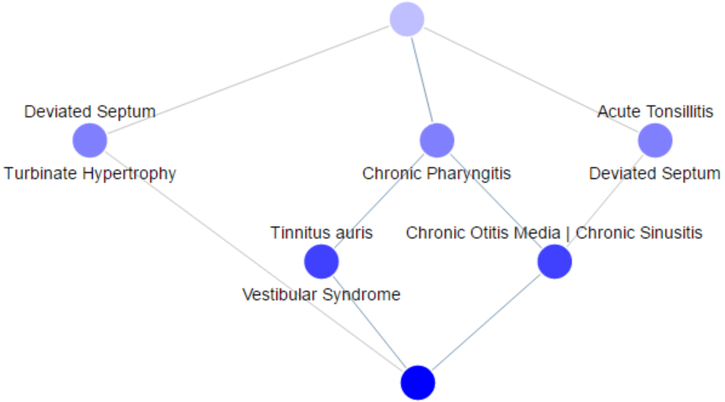

Fig. 1 - Interpretation

- Fig. 1 shows the conceptual scale for the relation between primary and secondary diagnostics which occurred for female and male patients

Problem:

- no straightforward representation for the triadic setting, for instance: investigate gender of patients against principal and secondary diagnostics

Solution:

- model the same data subset in a graph database and visualize the dat graph showing correlations

Using *Neo4j*

- in order to have a different perspective, the data can be processed and stored in a graph database
- we visualize data graphs and analyze connections between the data using *Neo4j*
- to make graphs easy to read, one can choose to exclude or include certain vertices (relations)
- in the following examples we chose to visualize certain relationships between diagnostics, symptoms and patients

Fig. 2. Correlations between gender, principal and secondary diagnostic

Fig. 2 - Interpretation

- Patient (green) has Principal Diagnostics (red) and arrow is labelled with *hasPD*
- Principal Diagnostics (red) might generate Secondary Diagnostics (blue) and arrow is labelled *hasSD*

Advantages of graph representation:

- the quantitative information of the clusters can be easily observed in the graphs, while it is not straightforward in the concept lattice

Observations in Fig. 2:

- principal diagnostics *Deviated Septum* occurred primarily in female patients
- patients diagnosed with *Chronic Sinusitis* as the principal diagnostic are very likely to have *Deviated Septum* or *Chronic Pharyngitis* as a secondary diagnostics

Fig. 3. Correlations between gender, principal symptoms and secondary symptoms

Fig. 3 - Interpretation

- here, we analyze the symptoms that led to the diagnostics
- an interesting fact that stands out is that certain symptoms only occurred for male patients, while others only for female patients (for the analyzed data)
- observations like this can lead, over a lasting analyze (that involves large datasets), to the formulation of hypothesis which can then be researched in more details by medical staff

Fig. 4. *Deviated Septum* (red) as principal diagnostic in relation to all symptoms and secondary diagnostics (blue)

Fig. 4 and Fig. 5 - Interpretation

- Fig. 4 shows all the symptoms (principal - yellow - and secondary - violet) and secondary diagnostics in relation to the principal diagnostics *Deviated Septum*
- it offers an overview of other diagnostics that occurred for patients with similar symptoms
- it offers an overview of all the symptoms that occurred for patients with *Deviated Septum*
- by eliminating certain vertices, in Fig. 5 we focus on *Chronic Sinusitis* as secondary diagnostic
- for instance, we observe that patients who were diagnosed with *Deviated Septum* had one of the following principal symptoms: *headache*, *rhinorrhea*, *nasal obstruction*, *tinnitus*

Fig. 5. *Deviated Septum* (red) as principal diagnostic and *Chronic Sinusitis* (blue) as secondary diagnostic in relation to corresponding symptoms

Conclusions and Future Work

- this is a first approach to combine the analytical capabilities of FCA with graph data analysis
- we have exemplified the approach on medical data
- by analyzing the obtained results, doctors can:
  - visualize correlations between their patients
  - coordinate treatment
  - analyze the differences between patients and potential disease progressions
- we intend to prove that the qualitative concept data analysis capabilities of FCA can be combined with a more quantitative approach in order to investigate local data correlations → conceptual landscapes can be enriched with new exploration capabilities based on data graphs