Identification of co-occurring diseases using ontological data mining techniques
Mihail Popescu
Health Management and Informatics Dept., University of Missouri, Columbia, MO

Abstract
In this work we investigate the identification of co-occurring diseases in a patient database using two ontological data mining techniques: fuzzy clustering and fuzzy co-clustering. We present preliminary results on a pilot data set of 107 patients.

Problem description
Thorpe [1] mentions that in 2002 about 40% of the Medicare beneficiaries were treated simultaneously for 5 or more conditions, accounting for about 65% of overall spending. The question is what are the groups of conditions in a given hospital? By identification of the co-occurring diseases in a hospital database one can stream-line the care and reduce its cost.

Ontological clustering
Ontological methods are used in conjunction with classical data mining algorithms (clustering, co-clustering) in order to avoid problems with diagnose matching. In this work we used ICD9 for describing the diagnoses. We are investigating several ontological data mining methods in order to assess their strengths and weaknesses. We use a pilot dataset that consisted in 107 patients each described with a number of 5 to 40 ICD9 diagnoses terms.

The first approach [2] consists in computing the semantic similarity between patients based on their ICD9 diagnoses followed by the clustering of the resulting similarity matrix [3]. Each identified patient cluster is then scanned for the most frequent diagnoses (see figure 1).

Ontological co-clustering
The second approach consists in describing the patients in term of their ICD9 codes in a diagnosis space $\Delta=\{\text{ICD9}_1, ..., \text{ICD9}_n\}$ where $n$ is the number of diagnoses of interest ($n=124$ here). Given a patient $P_i$ and its ICD9 diagnose set $\{\text{ICD9}_{i1}, ..., \text{ICD9}_{im}\}$, one can describe $P_i\in\Delta$ as $P_i=(\max_{x=1...n}[s(\text{ICD9}_j,\text{ICD9}_{ij})])...\max_{x=1...n}[s(\text{ICD9}_j,\text{ICD9}_{is})])$, where $s(x,y)$ is the similarity [2] between the ICD9 codes $x$ and $y$.

The advantage of this representation compared to the regular vector space approach is that $n<<9000$ (~total # of ICD9 codes).

As we can see in figure 2, the ontological co-clustering [4] identified roughly the same patient groups with somewhat similar diagnoses codes.

Conclusion
We are currently investigating two methods for identifying co-occurring diseases in patient databases. The clustering of the patient similarity matrix can include other patient information such as demographics and lab-tests but requires sampling for more than 20,000 patients. The co-clustering approach is more scalable but the resulting ICD9 code are less precise.

References