Taxonomy-based dissimilarity measures for profile identification in medical data

Roxana Dogaru, Flavia Micota, Daniela Zaharie
Department of Computer Science
West University of Timişoara
blvd. Vasile Pârvan, 4, 300223 Timişoara
Email: {rdogaru|zflavia|dzaharie@info.uvt.ro}

***Abstract*—The lists of diagnostic codes which are usually recorded in the hospitals for health management and/or costs reimbursement purposes can represent a useful source of information in the analysis of the (dis)similarity between different patients, as long as appropriate measures exist to estimate this (dis)similarity. The aim of this paper is to analyze various measures obtained by using different ways of computing the information content corresponding to entities in a taxonomy and by aggregating different types of measures. The discriminative power of these measures is evaluated by analyzing their ability to explain existing groups in data. A case study based on medical records containing lists of ICD (International Classification of Diseases) codes is presented and the proposed dissimilarity measures are used to identify prototypes in groups of patients.**

I. Introduction

The electronic medical records, usually filled in hospitals for clinical and/or administrative purposes, represent a wealthy source of information which could be exploited in order to develop clinical prediction or decision support tools. Besides unstructured information, the medical records may contain structured information specified using existing nomenclatures and taxonomies which systematize the medical concepts (e.g ICD - International Classification of Diseases, SNOMED CT - Systematized Nomenclature of Medicine - Clinical Terms etc.). An example of such structured information is represented by the lists of ICD codes corresponding to the principal/secondary diagnostics and medical procedures which are recorded for each hospitalized patient. The aim of this paper is to investigate how such lists can be used in order to assess the similarity between different patients, to group patients with similar pathologies or to identify patient profiles. For any of these tasks a critical element is the identification of an appropriate (dis)similarity measure. Currently there exist a plethora of (dis)similarity measures designed both for standard types of data (e.g. binary, numerical) as well as for concepts belonging to medical ontologies [11], [9]. In this context several questions arise: (i) how can be extended existing semantic dissimilarity measures to lists of taxonomic codes? (ii) how are related these measures? (iii) is the aggregation of several measures beneficial in the context of discriminating groups in data? Aiming to answer these questions, several dissimilarity measures are analyzed and a case study based on data collected at the Obstetrics and Neonatology wards of a hospital is conducted.

The list of ICD codes have been previously used in several data mining tasks but in most cases they have been transformed in binary or numerical vectors. For instance in [4] the lists of ICD codes were aggregated in several groups and translated into occurrence vectors. In [10] a TF-IDF representation is constructed based on ICD codes associated to each patient and a cosine similarity is used in order to identify clusters in the set of patients. In [6] the codes are intrepreted as features and the structure of the ICD taxonomy is used in order to design stable feature selection algorithms.

The main difference between such approaches and that presented in the current paper is the fact that here it is analyzed directly the list of codes without translating them in high-dimensional occurrence/frequency vectors, but by exploiting the structure of the ICD taxonomy. The particularities of this taxonomy are shortly presented in the second section. Section III presents the proposed taxonomy-based dissimilarity measures while the relationship between them is analyzed in the fourth section, where results on their equivalence degree are presented. The results of an experimental study conducted in the case of data collected from an Obstetrics and Gynaecology hospital are summarized in Section V and conclusions are provided in the last section.

II. International Classification of Diseases

The International Classification of Diseases (ICD) is a taxonomy of medical diagnostics and procedures and is used to standardize the descriptions of health problems. Its current version is ICD-10[1]. This hierarchical structure consists of four main levels corresponding to chapters, groups, sections and codes and each branch in the hierarchy has the same length. A sample from the ICD-10 taxonomy is illustrated in Figure 1. Each diagnostic has an associated code and related pathologies are "closer" in the ICD tree than non-related pathologies. The ICD codes, as other taxonomic codes, are hierarchical, meaning that they carry information related to the chapter, group or section to which they belong. For instance the difference between two codes belonging to the same section (e.g. O601 corresponding to "Preterm spontaneous labour with preterm delivery" and O603 corresponding to "Preterm delivery without spontaneous labour") is smaller than the difference between two codes belonging to different chapters (e.g. O601 and H101 corresponding to "Acute atopic conjunctivitis").

The health record of a hospitalized patient usually contains a list of ICD codes corresponding to the primary and secondary diagnostics and to the medical procedures applied to that patient. For instance a typical list of codes corresponding to a

[1] http://www.who.int/classifications/icd/en/

women hospitalized in a Obstetrics ward to give birth to a child could be [O731, Z370, Z390, Z391, Z392]. By using such lists of taxonomic codes the (dis)similarity between various cases can be assessed and patient profiles can be identified.

III. Dissimilarity Measures for Lists of Diagnostic Codes

Let A and B be two lists of taxonomic codes. The simplest way of computing the dissimilarity between A and B is to count how many elements are present only in one of these lists, as it is done for instance in the case of the so-called Dice dissimilarity:

$$d_{Dice}(A,B) = \frac{|A \backslash B| + |B \backslash A|}{|A| + |B|} = 1 - \frac{2|A \cap B|}{|A| + |B|} \quad (1)$$

Such a dissimilarity measure is appropriate in the case when it is of interest to check only if the elements of A and B are identical or not. In the case of lists of diagnostic codes there are several degrees at which two codes are different, e.g. they can be different at the chapter level or only at the group/ section/ code level. Therefore the simple comparison between codes which leads only to binary outcomes (identical or different codes) should be replaced with the computation of a dissimilarity based on the degree of closeness between the codes in the taxonomic structure.

The main idea in computing the similarity between two concepts belonging to a taxonomy or in a more general context to an ontology is to estimate the difference between the amount of common and specific aspects [9], [12]. The commonality amount is usually expressed by the information content of the *least common ancestor* of the two concepts in the taxonomy. For instance in the case of the codes O601 and O603 the least common ancestor is O60-O75 (section "Complications of labour and delivery") while for O601 and H101 the least common ancestor is the root of the taxonomy, meaning that they do not have common elements (according to ICD-10 taxonomy). Most taxonomy-based or semantic dissimilarity measures are based on the combination between the information content (IC) of the least common ancestor and the information content of the compared concepts or codes. One way of combining the information content of the common and specific parts of concepts is based on the same structure as in the Dice dissimilarity and is described in Eq. (2) for two codes C_1, C_2 and their least common ancestor, $lca(C_1, C_2)$.

$$\delta(C_1, C_2) = 1 - \frac{2IC(lca(C_1, C_2))}{IC(C_1) + IC(C_2)} \quad (2)$$

Various measures differ in the way the information content is estimated, e.g. based on edge counting, on corpus-based probabilities or on the taxonomy content. In this paper a variant of each type is analyzed.

The Wu-Palmer dissimilarity (δ_{WP}) proposed in [13] uses the depth of the corresponding node in the taxonomy to estimate the information content, i.e. $IC_{WP}(C) = \text{depth}(C)$. In the case of the ICD-10 taxonomy the node corresponding to the taxonomy root has a depth equal to 0, the nodes corresponding to chapters have the depth equal to 1, those corresponding to groups have the depth equal to 2, etc. A disadvantage of edge-counting measures is the fact that the information content of all concepts on the same level is the same, disregarding the degree of complexity of the structure rooted in it.

A corpus based measure uses the probabilities estimated using the relative frequency of the concepts in a data corpus to compute the information content, i.e. $IC_P(C) = -\log(Prob(C))$. This measure has been first proposed by Lin [8]. The main disadvantage of this measure is that it is highly dependent on the available corpus.

A more recent measure has been proposed by Sanchez et al. in [12] and tries to overcome the disadvantages of the previous measures. Its main idea is to use the ratio between the generality of a concept (estimated by the number of leaves in the sub-tree rooted in the concept, $|\mathcal{L}(C)|$) and its concreteness (which is related to the number of taxonomical ancestors including itself, $|\mathcal{A}(C)|$), as is described in Eq.(3) where $|\mathcal{L}(T)|$ denotes the number of leaves of the entire taxonomy.

$$IC_S(C) = -\log\left(\frac{|\mathcal{L}(C)|/|\mathcal{A}(C)| + 1}{|\mathcal{L}(T)| + 1}\right) \quad (3)$$

By combining these variants to estimate the information content with the dissimilarity computing rule described in Eq. (2) one obtains three taxonomy-based dissimilarity measures denoted in the following by δ_{WP}, δ_P and δ_S.

In order to extend these dissimilarities to lists of taxonomic codes one can generalize the set-based dissimilarity described in Eq. (1) in order to take into account the dissimilarity degree between codes as is described in Eq. (4).

$$d_D(A,B) = \frac{\sum_{a \in A} \min_{b \in B} \delta(a,b) + \sum_{b \in B} \min_{a \in A} \delta(b,a)}{|A| + |B|} \quad (4)$$

As the dissimilarity measures between codes are based on different types of information it makes sense to analyze the opportunity to combine them in order to obtain dissimilarity measures with potentially better discriminative power. It is expected that the benefits of combining several measures is influenced by their degree of equivalence or complementarity, i.e. their ability to capture same or different aspects. Therefore the next section is devoted to the analysis of the equivalence degree between the code-level dissimilarity measures δ_{WP}, δ_P and δ_S which rely on IC_{WP}, IC_P and IC_S, respectively.

IV. Equivalence degree between dissimilarity measures

When applied to data mining tasks, different dissimilarity measures usually lead to different results. However when the data mining task relies mainly on comparisons between dissimilarity values, as happens for instance in hierarchical agglomerative clustering, then different dissimilarity measures can lead to the same or very similar results. Such measures could be considered equivalent in some sense and knowing their degree of equivalence can be useful in constructing new measures by aggregating existing ones. The concept of equivalence between resemblance measures has been introduced in [2] and relies on the partial order induced by the measures on the set of pairs of entities. More specifically, two dissimilarity

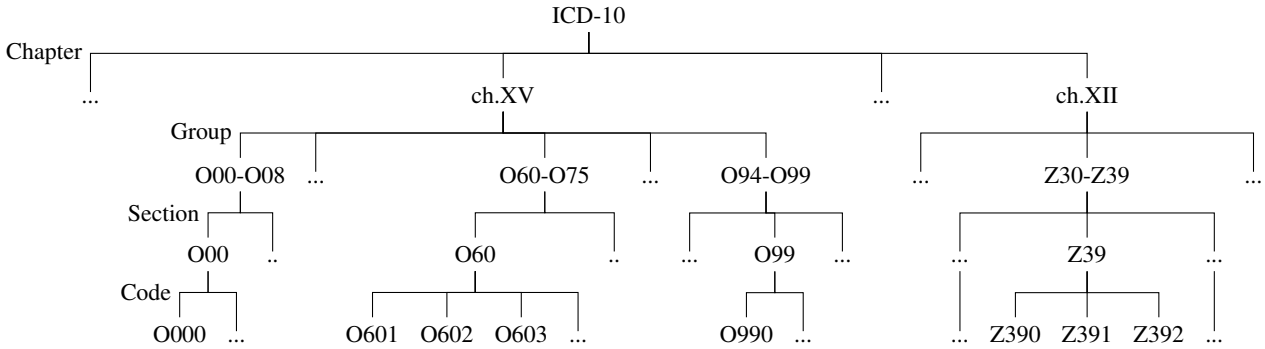

Fig. 1. Fragment of the ICD-10 taxonomy

measures d_1 and d_2 are considered equivalent if for any 4-tuple (a, b, c, d) of entities the following statements are true:
(i) $d_1(a, b) < d_1(c, d)$ if and only if $d_2(a, b) < d_2(c, d)$;
(ii) $d_1(a, b) = d_1(c, d)$ if and only if $d_2(a, b) = d_2(c, d)$.

This property has been analyzed for measures corresponding to binary and numerical data [7] but, at our best knowledge, there are no reported results for taxonomy-based measures. In the following is analyzed under which conditions the dissimilarities δ_{WP}, δ_P and δ_S are equivalent.

Let (C_1, C_2, C_3, C_4) be an arbitrary 4-tuple of codes but having all the same depth in the taxonomy (in the context of ICD-10 this means that they are full 4-characters codes). Thus $\text{depth}(C_i) = \text{depth}(T)$ for $i = \overline{1,4}$ and it follows that $\delta_{WP}(C_1, C_2) < \delta_{WP}(C_3, C_4)$ is equivalent to $\text{depth}(lca(C_1, C_2)) > \text{depth}(lca(C_3, C_4))$.

Equivalence between δ_{WP} and δ_P. Supposing that the corpus used to compute δ_P contains uniformly distributed codes (i.e. $Prob(C_i) = p$) then it follows that $\delta_P(C_1, C_2) < \delta_P(C_3, C_4)$ if and only if $Prob(lca(C_1, C_2)) < Prob(lca(C_3, C_4))$. If the probability associated to a node in the taxonomy is inverse proportional to its depth then this last inequality is equivalent to $\text{depth}(lca(C_1, C_2)) > \text{depth}(lca(C_3, C_4))$ and consequently to $\delta_{WP}(C_1, C_2) < \delta_{WP}(C_3, C_4)$. If the probabilities corresponding to nodes having the same depth are also equal then the condition on equal distances is also satisfied and it follows that δ_{WP} and δ_P are equivalent. It should be remarked that the assumptions made on the probabilities estimated from the corpus data are rather strong.

Equivalence between δ_{WP} and δ_S. Suppose now that for any two nodes c_i and c_j from the taxonomy the following assumptions are satisfied:
(i) $\text{depth}(c_i) > \text{depth}(c_j)$ if and only if $|\mathcal{L}(c_i)| < |\mathcal{L}(c_j)|$;
(ii) $\text{depth}(c_i) = \text{depth}(c_j)$ if and only if $|\mathcal{L}(c_i)| = |\mathcal{L}(c_j)|$.
This means that two nodes are on the same level if and only if the subtrees rooted in them have the same number of leaves and as the level is closer to the taxonomy root the number of leaves is higher. For such a "balanced" taxonomy it can be proven that δ_{WP} is equivalent with δ_S. Indeed, one can prove by contradiction that $\text{depth}(lca(C_1, C_2)) > \text{depth}(lca(C_3, C_4))$ if and only if $|\mathcal{L}(lca(C_1, C_2))|/(\text{depth}(lca(C_1, C_2)) + 1) < |\mathcal{L}(lca(C_3, C_4))|/(\text{depth}(lca(C_3, C_4)) + 1)$ which, under the assumption that all codes have the same depth, is equivalent with $\delta_S(C_1, C_2) < \delta_S(C_3, C_4)$. On the other hand the assumptions on the taxonomy ensures that $\text{depth}(lca(C_1, C_2)) = \text{depth}(lca(C_3, C_4))$ if and only if $|\mathcal{L}(lca(C_1, C_2))|/(\text{depth}(lca(C_1, C_2)) + 1) = |\mathcal{L}(lca(C_3, C_4))|/(\text{depth}(lca(C_3, C_4)) + 1)$.

The above sufficient assumptions for the equivalence of δ_{WP}, δ_P and δ_S are rather restrictive and in practice are rarely satisfied. Therefore in the following is analyzed a relaxed version of equivalence which can be expressed by the equivalence degree between measures defined by the ratio of 4-tuples for which the two dissimilarities induce the same partial order over pairs of codes[14]:

$$E(d_1, d_2) = \frac{1}{\text{card}S} \sum_{(a,b,c,d) \in S} \Delta_{abcd} \tag{5}$$

$$\Delta_{abcd} = \begin{cases} 1 & \text{if } (d_1(a,b) - d_1(c,d))(d_2(a,b) - d_2(c,d)) > 0 \\ & \text{or } d_1(a,b) = d_1(c,d) \text{ and } d_2(a,b) = d_2(c,d) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

When $E(d_1, d_2) = 1$ then d_1 and d_2 are equivalent while when $E(d_1, d_2) = 0$ there is a full discordance between the dissimilarities. It is expected that as $E(d_1, d_2)$ is higher the results of a data mining method which uses the order induced by d_1 and d_2 are closer.

Let analyze now the equivalence degree between an aggregated dissimilarity and the individual dissimilarities. Let d_1 and d_2 be two dissimilarities and $d_{min} = \min\{d_1, d_2\}$, $d_{max} = \max\{d_1, d_2\}$, $d_{avg} = (d_1 + d_2)/2$ some aggregated variants. Let (a, b, c, d) be a 4-tuple such that $(d_1(a, b) - d_1(c, d))(d_2(a, b) - d_2(c, d)) > 0$. It is easy to check that $(d_{min}(a, b) - d_{min}(c, d))(d_i(a, b) - d_i(c, d)) > 0$ for any $i \in \{1, 2\}$. Similarly if $d_1(a, b) = d_1(c, d)$ and $d_2(a, b) = d_2(c, d)$ then $d_{min}(a, b) = d_{min}(c, d)$. Thus any 4-tuple counted in the computation of $E(d_1, d_2)$ is also counted in the computation of $E(d_{min}, d_i)$ meaning that $E(d_{min}, d_i) \geq E(d_1, d_2)$. In a similar manner one can prove that $E(d_{max}, d_i) \geq E(d_1, d_2)$ and $E(d_{avg}, d_i) \geq E(d_1, d_2)$. This means that by aggregating two dissimilarities it is obtained a measure having an equivalence degree with each of the component dissimilarities higher than the degree of equivalence between them.

Table I presents the equivalence degree corresponding to dissimilarities over lists of codes obtained by combining the dissimilarity at list level d_D given in Eq. (4) with the code-level measures (δ_{WP}, δ_P and δ_S). The reported values have been computed using a set of 1169 lists of ICD-10 codes. As it is expected the equivalence degree between a random

Equivalence degree		
$E(\delta_{WP}, \delta_P) = 0.916$	$E(\delta_{WP}, \delta_S) = 0.920$	$E(\delta_P, \delta_S) = 0.854$
$E(\delta_{WP}, R) = 0.499$	$E(\delta_P, R) = 0.499$	$E(\delta_S, R) = 0.500$

TABLE I. EQUIVALENCE DEGREES BETWEEN δ_{WP}, δ_P, δ_S AND A RANDOM DISTANCE R

dissimilarity and each of the analyzed variants is very close to 0.5. On the other hand, the equivalence degree for each pair of non-random dissimilarities is rather high, particularly for δ_{WP} and δ_S. This can be explained by the fact that only 1.5% of the internal nodes of the ICD-10 taxonomy does not satisfy the property stating that a node with a larger depth has a smaller number of leaves in its corresponding subtree, i.e. $\text{depth}(c_i) > \text{depth}(c_j)$ if and only if $|\mathcal{L}(c_i)| < |\mathcal{L}(c_j)|$.

V. A CASE STUDY

The case study is based on data collected at Obstetrics and Neonatology wards of a hospital and consists of pairs of list of ICD-10 codes corresponding to mothers and their newborn(s). As the aim was to identify relationships between mother and newborn pathologies the analysis was focused on the estimation of the similarity between patients belonging to some groups and on the identification of profiles which are specific to these groups. In this context are considered two criteria to stratify the data: (i) the principal diagnostic of the newborn; (ii) the mother age.

The partitions constructed using these criteria have been used as ground truth to assess the discriminative ability of the dissimilarity measures. The idea of assessing the quality of a (dis)similarity measure by using clustering validity indices has been recently proposed in [3]. The same idea is used here but, based on the results of some comparative studies on cluster quality assessment [5], [1], two cluster validity indices are combined: C-index and Silhouette. C-index takes values in $[0, 1]$ and smaller values suggest better agreement between the existing partition and the dissimilarity value. On the other hand, the Silhouette index takes values in $[-1, 1]$ and higher values suggest a better quality. Thus the combined value which has been used is:

$$CS(\mathcal{P}) = \frac{1}{2}(\text{C-index}(\mathcal{P}) + (1 - \text{Silhouette}(\mathcal{P}))/2). \quad (7)$$

A. Analyzed dissimilarity measures

Several groups of dissimilarity measures for lists of taxonomic codes have been included in this analysis.

A first group consists of variants obtained by combining some well known list-level similarities (e.g. Dice, Jaccard, Cos, Overlap) with a binary code-level measure (which returns 0 when the codes are identical and 1 in all the other cases). All of these are traditional set-based measures (first group of four measures in Table II).

A second group contains three measures obtained by combining the taxonomy-based dissimilarities presented in Section III (δ_{WP}, δ_P, δ_S) with the set-level dissimilarity described in Eq. (4). These measures are based on the information content computed using Eq. (2) which is obviously related to the Dice dissimilarity. By using different ways of computing the information content one can obtain other measures. In this analysis are included two other measures obtained by using a formula similar to the Jaccard similarity to compute the information content. These variants correspond to the third group of measures in Table II.

The last group is represented by several aggregated measures obtained by combining the code-level dissimilarities or by using both extrinsic and intrinsic information in the computation of IC. The last variant included in the fourth group of Table II is based on the computation of the information content by using both the corpus-based probability and the taxonomy structure.

B. Results and discussion

Table II presents the values of the combined cluster quality index (based on C-Index and Silhouette) computed for four data partitions. The first and the third data groupings (denoted by Set 1 and Set 3) correspond to partitioning of the newborn records based on the newborn principal diagnostic and on the mother age, respectively. The second and the fourth data groupings (denoted by Set 2 and Set 4) correspond to the partitioning of the mother records also by using the newborn principal diagnostic and the mother age, respectively.

By analyzing the results reported in Table II the following remarks can be made:

(i) The measures using taxonomy based code-level dissimilarities (last three groups) explain better all data partitions than the traditional measures (first group), as the corresponding values of CS are smaller. Thus, the usage of the ICD taxonomy structure improves the discriminative ability of the corresponding measures.

(ii) The results corresponding to the second and the third groups illustrate that both the variant of computing the information content and the way the code-level similarity is computed have an influence on the overall behavior of a measure. For all analyzed datasets the best agreement between the measure and the existing partition in data is observed in the case of the code-level measure which uses the taxonomy structure (δ_S) and the combination of the IC-values described in Eq. (2).

(iii) The aggregation of different measures does not provide an improvement of the discriminative ability with respect to the individual measures. This can be partly explained by the high equivalence degree between the component measures (as illustrated in Table I). In this fourth group the best behavior corresponds to the $\min$ aggregation variant, the results being identical with those obtained for D_S. This result would suggest that in most cases D_S leads to the smallest value when compared with D_{WP} and D_P. Moreover, the results obtained for the $\max$ and median suggest that for most pairs (L_1, L_2) of lists of codes the following inequality is satisfied: $D_S(L_1, L_2) \leq D_{WP}(L_1, L_2) \leq D_P(L_1, L_2)$. Finally, combining the corpus-based probability with the intrinsic characteristics of the taxonomy in the computation of the information content ($D_{C(P,S)}$) leads to a slight improvement with respect to D_P but no improvement is obtained with respect to D_S.

Based on these comparative results the D_S dissimilarity is further used to identify prototypes of several groups of data. As prototype for a group of code lists G is considered the

Dissimilarity measures				Aggregated C-Index and Silhouette values			
Notation	**List level**	**Code level**	**IC(C)**	**Set 1**	**Set 2**	**Set 3**	**Set 4**
Dice	$1-\frac{2\lvert A\cap B\rvert}{\lvert A\rvert+\lvert B\rvert}$	$0/1$	-	0.477	0.504	0.483	0.534
Jaccard	$1-\frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$	$0/1$	-	0.516	0.535	0.507	0.556
Cos	$1-\frac{\lvert A\cap B\rvert}{\sqrt{\lvert A\rvert\cdot\lvert B\rvert}}$	$0/1$	-	0.478	0.501	0.482	0.533
Overlap	$1-\frac{\lvert A\cap B\rvert}{\min\{\lvert A\rvert,\lvert B\rvert\}}$	$0/1$	-	0.545	0.518	0.498	0.534
D_{WP}	d_D(Eq.4)	$1-\frac{2IC(lca(C_1,C_2))}{IC(C_1)+IC(C_2)}$	depth(C)	0.422	0.462	0.464	0.488
D_P			$-\log(Prob(C))$	0.430	0.481	0.468	0.502
D_S			$-\log\left(\frac{\lvert\mathcal{L}(C)\rvert/\lvert\mathcal{A}(C)\rvert+1}{\lvert\mathcal{L}(T)\rvert+1}\right)$	**0.406**	**0.442**	**0.450**	**0.470**
J_P	d_D(Eq.4)	$1-\frac{IC(lca(C_1,C_2))}{IC(C_1)+IC(C_2)-IC(lca(C_1,C_2))}$	$-\log(Prob(C))$	0.444	0.492	0.474	0.513
J_S			$-\log\left(\frac{\lvert\mathcal{L}(C)\rvert/\lvert\mathcal{A}(C)\rvert+1}{\lvert\mathcal{L}(T)\rvert+1}\right)$	0.421	0.456	0.462	0.482
D_{max}	d_D(Eq.4)	$\max\{\delta_{WP},\delta_P,\delta_S\}$		0.430	0.482	0.468	0.503
D_{min}		$\min\{\delta_{WP},\delta_P,\delta_S\}$		**0.406**	**0.442**	**0.450**	**0.470**
D_{avg}		$(\delta_{WP}+\delta_P+\delta_S)/3$		0.418	0.462	0.461	0.488
D_{median}		median$\{\delta_{WP},\delta_P,\delta_S\}$		0.421	0.461	0.464	0.488
D_{WP+P}		$(\delta_{WP}+\delta_P)/2$		0.413	0.452	0.457	0.479
D_{P+S}		$(\delta_P+\delta_S)/2$		0.417	0.462	0.460	0.488
$D_{C(P,S)}$		$1-\frac{2IC(lca(C_1,C_2))}{IC(C_1)+IC(C_2)}$	$-\frac{1}{2}\log\left(Prob(C)\frac{\lvert\mathcal{L}(C)\rvert/\lvert\mathcal{A}(C)\rvert+1}{\lvert\mathcal{L}(T)\rvert+1}\right)$	0.411	0.451	0.456	0.479

TABLE II. COMPARISON BETWEEN DISSIMILARITY MEASURES FOR LISTS OF TAXONOMIC CODES USING THE AGGREGATION BETWEEN C-INDEX AND SILHOUETTE AS INDICATOR OF THE ABILITY OF THE MEASURE TO EXPLAIN EXISTING GROUPS IN THE DATA: SETS 1 AND 3 CORRESPOND TO PARTITIONS OF NEWBORNS RECORDS BASED ON THEIR PRINCIPAL DIAGNOSTIC AND ON THE MOTHER AGE, RESPECTIVELY; SETS 2 AND 4 CORRESPOND TO PARTITIONS OF MOTHERS RECORDS BASED ON NEWBORN PRINCIPAL DIAGNOSTIC AND ON THE MOTHER AGE, RESPECTIVELY. SMALLER VALUES OF THE INDICATOR SUGGEST BETTER DISCRIMINATIVE ABILITY.

element P^* having the property that the sum of dissimilarities with respect to all other elements in the group is minimal, i.e.

$$\sum_{P\in G} D_S(P^*,P) \leq \sum_{P\in G} D_S(P',P), \quad \text{for any } P'\in G.$$

Table III presents examples of prototypes obtained for groups of newborn records stratified by the mother age and for groups of mother records stratified by their newborn principal diagnostic class. The newborn diagnostic codes present in the analyzed data corresponds to various cases: newborns without a specific pathology (in this case the main diagnostic code corresponds to Z* class meaning that routine medical procedures have been applied), newborns with pathologies occuring in the perinatal period (corresponding to P* codes) and newborns with other types of pathologies. By analyzing these prototypes and the dissimilarities between them one can infer relationships between newborn health particularities and their mother age and/or health status. For instance looking at the dendrogram in Fig. 2 (left) one can see that the newborns health status is in some sense related to the mother age and more similar profiles are observed for closed groups of age (e.g. 13-17 years with 18-24 years and 25-30 with 31-35 years). The dendrogram on the right part of the same figure illustrate similarities between health profiles of mothers when they are grouped based on the pathologies of their newborns. Both dendrograms have been obtained by applying a single-link aglomerative clustering algorithm on the set of D_S values computed for the prototypes corresponding to the various groups.

VI. CONCLUSIONS

Starting from existing ontology-based similarity measures based on various ways of computing the information content of a concept are constructed several dissimilarities for lists of taxonomic codes (as are the lists of ICD codes present in the medical records of hospitalized patients). As each type of measure exploits different data or taxonomy information the possibility of constructing new measures by aggregating existing ones has been analyzed. Using the concept of equivalence degree between measures the relationship between measures based on different ways of estimating the information content (edge-counting, corpus-based probabilities, taxonomy structure) has been investigated. This investigation led to the conclusion that if the taxonomy is "balanced" and the distribution of codes in the data is almost uniform then the codel-level measures (δ_{WP},δ_P and δ_S) are almost equivalent. This relationship has been confirmed by the values of the empirically estimated equivalence degrees between the dissimilarities over lists of codes (D_{WP}, D_P and D_S).

In order to assess the quality of a dissimilarity measure it has been evaluated how well the measure can explain observed groups in data. This has been done using a combination of two known cluster validity indices (C-index and Silhouette)

Mother age	No.of cases	Prototype	Newborn diagnostic class	No.of cases	Prototype
13-17	57	[P599, Z232, Z246, Z298]	P05*(slow fetal growth)	56	[O342, O990, Z370, Z390, Z391, Z392]
18-24	490	[P599, Z232, Z246, Z298, Z380]	P07*(low birth weight)	85	[O347, O730, O990, Z370, Z390, Z391, Z392]
25-30	659	[P034, P599, P614, Z232, Z246, Z298, Z380]	P08*(high birth weight)	45	[O348, O713, O990, Z370, Z390, Z391, Z392]
31-35	372	[P034, P599, P613, Z232, Z246, Z298, Z380]	P12*(birth injury to scalp)	63	[O347, O713, Z370, Z390, Z391, Z392]
36-40	125	[P034, P599, Z232, Z246, Z298, Z383]	P20*(respiratory and cardiovascular disorders)	108	[O680, O990, Z370, Z390, Z391, Z392]
over 40	18	[P025, P614, Z232, Z246, Z298, Z380]	P36*(bacterial sepsis of newborn)	35	[O321, Z370, Z390, Z391, Z392]
			P50*(fetal blood loss)	253	[O347, O730, O990, Z370, Z390, Z391, Z392]
			P70*(disorders of newborn metabolism)	9	[O800, O990, Z370, Z390, Z391]
			P80*(hypothermia of newborn)	42	[O321, O990, Z370, Z390, Z391, Z392]
			P0*(P00-P04:complications of pregnancy)	182	[O730, Z370, Z390, Z391, Z392]
			Q*(congenital malformations)	43	[O342, O990, Z370, Z390, Z391, Z392]
			R*(other abnormal clinical findings)	9	[O334, O820, Z370, Z391, Z392]
			D18*(haemangioma)	36	[O347, O800, Z370, Z3900, Z391, Z392]
			Z*(routine medical procedures)	374	[O731, Z370, Z390, Z391, Z392]

TABLE III. PROTOTYPES ASSOCIATED TO NEWBORN GROUPS CORRESPONDING TO DIFFERENT AGE INTERVALS OF THE MOTHER (LEFT PANEL) AND TO MOTHERS GROUPS CONSTRUCTED BASED ON THE PRINCIPAL DIAGNOSTIC/ MEDICAL PROCEDURE CORRESPONDING TO THE NEWBORN.

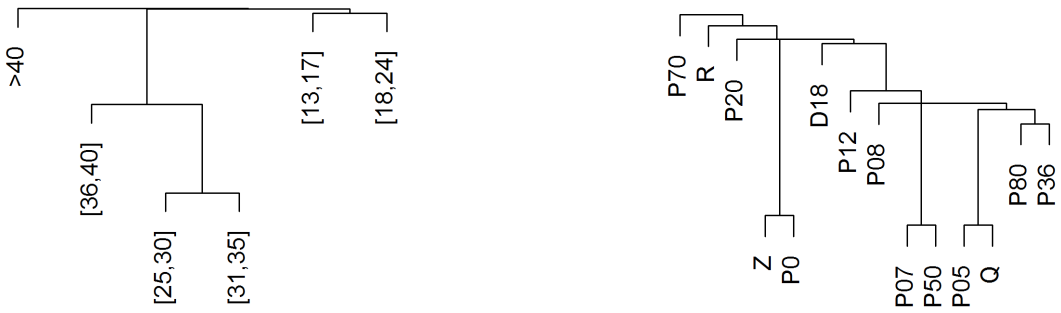

Fig. 2. Dendrograms, obtained by single-link agglomerative clustering using D_S, corresponding to the prototypes associated to two data partitions: partition of newborns records by mother age (left); partition of mother records by the newborn principal diagnostic (right).

as they have been identified by some previous studies to be effective. Based on this approach, 16 dissimilarity variants using four partitions of data have been compared. In all cases the best results have been obtained for the measure constructed starting from the semantic similarity proposed in [12]. On the other hand the aggregated measures have not provided advantages over the best performing component. This could be explained by the fact that there is a rather high degree of equivalence between the component measures. It is expected that an advantage is obtained by aggregation only in the case when the component measures uses complementary information.

On the other hand, in order to obtain more relevant patient profiles, the lists of diagnostic codes should be used in conjunction with other clinical information. This would require the aggregation of (dis)similarity measures over various domains (numerical, categorical, taxonomic codes or medical concepts belonging to specific ontologies), and it represents a line of further research.

REFERENCES

[1] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013.

[2] V. Batagelj and M. Bren. Comparing resemblance measures. *J. of Classification*, 12:73–90, 1995.

[3] R. Dogaru, F. Micota, and D. Zaharie. Searching for taxonomy-based similarity measures for medical data. In *Proceedings of the 7th Balkan Conference on Informatics Conference*, BCI '15, pages 27:1–27:8, New York, NY, USA, 2015. ACM.

[4] F. Doshi-Velez, Y. Ge, and I. Kohane. Comorbidity clusters in autism spectrum disorders: An electronic health record time-series analysis. *Pediatrics*, 133(1):e54–63, 2014.

[5] I. Gurrutxaga, J. Muguerza, O. Arbelaitz, J. M. Perez, and J. I. Martin. Towards a standard methodology to evaluate internal cluster validity indices. *Pattern Recognition Letters*, 32:505–515, 2011.

[6] I. Kamkar, S. K. Gupta, D. Phung, and S. Venkatesh. Stable feature selection for clinical prediction: Exploiting icd tree structure using tree-lasso. *J. of Biomedical Informatics*, 57(3):277–290, 2015.

[7] M.-J. Lesot and M. Rifqi. Order-based equivalence degrees for similarity and distance measures. In *International conference on Information Processing and Management of Uncertainty in knowledge-based systems*, pages 19–28, 2010.

[8] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, 1998.

[9] T. Pedersen, S. V. S. Pakhomov, S. Patwardhan, and C. G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *J. of Biomedical Informatics*, 40(3):288–299, 2007.

[10] F. Roque, P. Jensen, H. Schmock, M. Dalgaard, and M. e. a. Andreatta. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput Biol*, 7(8):e1002141, 2011.

[11] A. Sanchez and M. Batet. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *Journal of Biomedical Informatics*, 44(5):749–759, 2011.

[12] D. Sánchez, M. Batet, and D. Isern. Ontology-based information content computation. *Know.-Based Syst.*, 24(2):297–303, Mar. 2011.

[13] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138, 1994.

[14] D. Zighed, R. Abdesselam, and A. Hadgu. Topological comparisons of proximity measures. In *LNAI 7301*, pages 379–391, 2012.