Searching for Taxonomy-based Similarity Measures for Medical Data

Roxana Dogaru
West University of Timişoara
blvd. Vasile Pârvan, no. 4
300223 Timişoara, Romania
rdogaru@info.uvt.ro

Flavia Micota
West University of Timişoara
blvd. Vasile Pârvan, no. 4
300223 Timişoara, Romania
zflavia@info.uvt.ro

Daniela Zaharie
West University of Timişoara
blvd. Vasile Pârvan, no. 4
300223 Timişoara, Romania
dzaharie@info.uvt.ro

ABSTRACT

Identifying an appropriate measure of similarity between concepts is an important step in solving data mining tasks and designing decision support systems, being also a challenging problem because of the plethora of measures in the current use. When choosing a similarity measure the particularities of data to be processed and the availability of additional information should be taken into account. This paper addresses the problem of estimating the (dis)similarity between lists containing entities defined by a taxonomy and presents the results of a comparative analysis of several measures. The analysis is conducted in the context of processing lists of ICD-10 diagnostic codes. The set of analyzed similarities is obtained by combining several code-level and list-level measures. A weighted version of the measure based on edge counting is proposed and its ability to explain observed structures in data is estimated using some cluster validity indices. The comparative analysis involves also measures based on the information content estimated using either the taxonomy structure or a data corpus.

CCS Concepts

•**Information systems** → *Clustering and classification;*
•**Applied computing** → *Health care information systems;*
•**Computing methodologies** → *Genetic algorithms;*

Keywords

similarity measures, taxonomy, clustering validation indices, medical data

1. INTRODUCTION

(Dis)similarity measures represent key elements in assessing the similarity between concepts, in clustering tasks, in classification based on nearest neighbors and in other case based reasoning approaches. In the medical data analysis, such measures can be used in identifying similar cases, constructing patient profiles, comorbidity analysis, risk factors prediction and processing clinical documents [1, 8, 15].

Selecting a similarity measure is not an easy task as both the nature and the structure of the data to be processed as well as the available additional information (e.g. corpus, taxonomies, ontologies etc.) should be taken into account. Moreover, assessing the appropriateness of the measure to a given task and a dataset represents another challenge.

There exist various criteria used to evaluate the quality of a given data analysis task (e.g. classification, clustering etc.) supposing that a similarity measure has been already selected. In this paper we address a different problem: that of identifying a similarity measure which provide a high agreement with some ground truth observations. More specifically, we consider that there are some pre-established clusters in data (which we consider as ground truth) and we are looking for a similarity measure based on which a high value of a clustering quality criteria is obtained. Thus this is a reversed problem with respect to typical ones encountered in the clustering field: instead of searching for a partition when a similarity measure is given we are searching for a similarity measure based on some existing partition. This search is done in the context of processing patient data consisting of lists of diagnostics and medical procedures specified through ICD-10 codes. ICD (International Classification of Diseases[1]) denotes a hierarchical structure over the set of medical diagnostics and procedures based on which a coding system is defined, i.e. each diagnostic has an associated code and related diagnostics have similar codes.

In this context are analyzed several similarity measures between diagnostic codes, a weighted variant is proposed and its properties are theoretically studied. By combining the similarity measures between entities in a taxonomy with dissimilarities between lists (bags) of entities a set of eight measures were generated and their effectiveness in explaining existing data partitions has been tested for data sets collected at two Obstetrics and Gynaecology hospitals. The ground truth partition is constructed by stratifying the medical records of mothers based on various pathologies of their child/children (disorders related to length of gestation and fetal growth, congenital malformations etc.). The additional information which is used in the construction of the similarity measures is provided by the ICD-10 taxonomy.

The main contributions of this paper are:

(i) a weighted taxonomy-based similarity measure is proposed and the influence of the weights values on the effectiveness of the similarity in explaining observed

BCI '15 September 2-4, 2015, Craiova, Romania

DOI: 10.475/123_4

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

structures in the data is analyzed;

(ii) several combinations between code-level and set-level (dis)similarities are proposed and analyzed by using cluster validity indices as indicators of the (dis)similarity quality.

The rest of the paper is organized as follows. Section 2 presents several code-level similarities based on taxonomies and set-level dissimilarities which can be used for lists of codes. In section 3 is presented a weighted similarity measure which generalizes the measure based on counting edges in the taxonomy. Section 4 discusses the problem of assessing the ability of a similarity measure to explain the observed structure in data, section 5 presents the experimental results, and section 6 concludes the paper.

2. RELATED WORK

There are some recent works which address the problem of extracting knowledge from medical records containing clinical information specified through ICD codes.

In [3] lists of ICD codes were collected from medical records of patients with autistic disorders, aggregated in several groups and translated into occurrence vectors. A hierarchical clustering was applied on occurrence vectors and some medical trajectories have been obtained. In [9] a TF-IDF representation is constructed based on ICD-10 codes associated to each patient and a cosine similarity is used in order to identify clusters in the set of patients. The main difference between these approaches and that presented in the current paper is the fact that we analyzed directly the list of codes without translating them in high-dimensional occurrence/frequency vectors.

A different approach is presented in [6] where the structure of the ICD-10 tree is exploited in order to build a prediction model by using stable feature selection algorithms (e.g. Tree-Lasso). However the aim of [6] is to learn a regularized classification model without explicitly defining a similarity measure.

On the other hand, the idea of using weights in similarity measures has been used before in various contexts, particularly when the weights are associated to features. For instance, the authors of [1] use a weighted similarity measure in the context of processing ICF (International Classification of Functioning, Disability, and Health [2]) codes. However the setting of weights is only marginally discussed being mentioned only that larger weights should be assigned to terms corresponding to more specific ICF categories. A similarity between ICD codes is also used in [14] but it does not use weights being based on edges counting.

The problem of tuning or learning from data the values of weights associated to features is another recurrent topic in data mining. In [17] is presented an approach to learn the weights associated to terms appearing in documents in order to construct vectorial representations of documents (instead of TF-IDF representation). The proposed learning framework is based on a regularized loss function computed using reference similarity pairs. We use instead known partitions of data and the similarity performance is evaluated by cluster quality indices.

[2]http://www.asha.org/slp/icf/

3. TAXONOMY BASED SIMILARITY MEASURES

3.1 Similarity between entities in a taxonomy

Taxonomies are hierarchical classifications based on "is-a" relationships between the entities associated to nodes placed on the same branch in the hierarchy. There are taxonomies corresponding to various domains and some of these provide specific encodings corresponding to the entities/concepts placed on nodes which can be used to easily locate any entity in the hierarchical structure. Examples of such taxonomies are ICD for the medical field, MSC[3] (Mathematics Subject Classification) for the mathematical field, ACM[4] (Computing Classification System) for the computer science field etc. An excerpt from the ICD-10 classification is presented in Figure 1.

The taxonomies can be used to estimate the similarity between entities corresponding to their nodes. There are two main classes of taxonomy based similarity measures: (i) edge-counting measures which uses the relative position of the entities in the taxonomy; (ii) information-content based measures which exploit both the co-location of entities and additional information on the amount of information carried by the concepts incorporated in the taxonomy. Disregarding the class to which they belong, all measures are based on estimating the discrepancy between commonality and specificity of entities.

In most taxonomy based measures, a key element in the estimation of the commonality between two entities is represented by the most specific ancestor shared by the two entities. In the following, the most specific ancestor of two concepts C_1 and C_2 is denoted by $lca(C_1, C_2)$ (*least common ancestor*) and the taxonomy based similarity measures used in this study are reviewed. For an overview of similarities based on ontologies see for instance [11].

One of the most popular edge-counting measure is that proposed by Wu and Palmer [16] and defined in Eq. (1) where *depth* denotes the number of edges from the hierarchy root to the node.

$$s_{WP}(C_1, C_2) = \frac{2 \cdot \text{depth}(lca(C_1, C_2))}{\text{depth}(C_1) + \text{depth}(C_2)} \quad (1)$$

The measures based on information content have a similar structure but instead of using the depth in the taxonomy they use an estimation of the information carried by the concept:

$$s_{IC}(C_1, C_2) = \frac{2 \cdot IC(lca(C_1, C_2))}{IC(C_1) + IC(C_2)} \quad (2)$$

Different measures are characterized by different ways of estimating the information content, using either a data corpus or the intrinsic structure of the taxonomy. Lin [7] estimates the information content by a probability computed based on a data corpus, i.e. $IC(C) = -\log(P(C))$, where $P(C)$ is estimated based on the frequency of C in the data corpus. The Lin similarity (s_{Lin}) is defined by:

$$s_{Lin}(C_1, C_2) = \frac{2 \cdot \log P(lca(C_1, C_2))}{\log(P(C_1)) + \log(P(C_2))} \quad (3)$$

[3]http://www.ams.org/mathscinet/msc/

[4]http://www.acm.org/about/class/

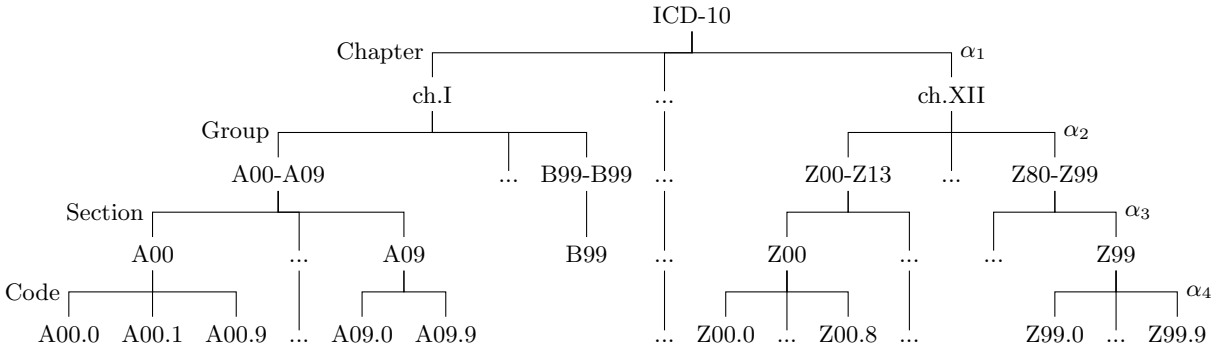

Figure 1: Fragment of the ICD-10 taxonomy

A variant which uses only the taxonomy structure has been proposed in [12] where the information content of a concept is expressed as the ratio between a measure of its generality (related to the number of ancestors) and a measure of its concreteness (related to the number of descendants). The specific estimation of the information content proposed in [12] is described in Eq. (4) where $|\text{leaves}(C)|$ denotes the number of leaves of the sub-tree having the node corresponding to C as root, and $|\text{ancestors}(C)|$ denotes the number of ancestors of C including itself. The similarity using this intrinsic estimation of the information content is denoted in the following by s_{SBI} (Sanchez-Batet-Isern similarity).

$$IC(C) = -\log\left(\frac{|\text{leaves}(C)|/|\text{ancestors}(C)|+1}{|\text{tree leaves}|+1}\right) \quad (4)$$

All the above measures take values in $[0,1]$ (the similarity of identical concepts is always 1, while the similarity of concepts having the root as least common ancestor is always 0). Examples of similarity values for examples of elements of the ICD-10 taxonomy are presented in Table 1. Thus it is easy to construct dissimilarity measures by just subtracting the similarity value from 1. In the rest of this paper the dissimilarity measures constructed based on these similarities are denoted by: $\delta_{WP} = 1 - s_{WP}$, $\delta_{Lin} = 1 - s_{Lin}$ and $\delta_{SBI} = 1 - s_{SBI}$.

3.2 Dissimilarity between sets of entities

Usually, the elements to be processed are not individual elements of a taxonomy but sets of such entities. For instance the medical record of a patient contains a list of diagnostic and procedure codes, a computer science paper may contain several subject categories etc. Thus estimating the dissimilarity between two patient profiles requires the extension of the similarities, as those presented in the previous subsection, to sets of taxonomic entities. If the used dissimilarity between diagnostic codes is a binary one (e.g. it provides 0 only if the codes are equal and 1 in all the other cases), then any set based dissimilarity (e.g. simple overlap, Dice, Jaccard, Tanimoto etc.) could be used. However, in order to preserve the discriminative information provided by the values of the dissimilarity between individual entities they should be involved in the computation of the dissimilarity between lists of entities. A natural idea is to use distances defined for sets of elements from metric spaces, as is the classical Hausdorff distance. For a taxonomy-based dissimilarity δ and two finite sets of entities, A and B, the Hausdorff dissimilarity is defined as in Eq. (5).

$$d_H(A,B) = \max\{\max_{a\in A}\min_{b\in B}\delta(a,b), \max_{b\in B}\min_{a\in A}\delta(b,a)\} \quad (5)$$

It is known that if δ is a distance then d_H is also a distance. On the other hand, it is easy to see that the set of possible values of d_H coincides with the set of values taken by δ. As it is considered that the discriminant capacity of distance functions is related to the number of distinct values which it can take [4], we propose a variant based on the concept of the Dice dissimilarity between sets but involving not necessarily binary dissimilarity between entities (Eq. (6)).

$$d_D(A,B) = \frac{\sum_{a\in A}\min_{b\in B}\delta(a,b) + \sum_{b\in B}\min_{a\in A}\delta(b,a)}{|A|+|B|} \quad (6)$$

In the experimental analysis both these measures are combined with all variants of code-level dissimilarities.

4. A WEIGHTED TAXONOMY BASED SIMILARITY

Starting from the fact that different levels in a taxonomy correspond to different types of correlations between the entities associated to those levels we propose to assign weights to levels. In this way each component of a taxonomic code could contribute in a different way to the code-level similarity. In this section we introduce a weighted code-level similarity and analyze its properties.

4.1 Definition, examples and properties

Let us consider two entities, C_1 and C_2, from a hierarchical taxonomy on which the levels are numbered starting with 0 for the root node and ending with the level k of leaves. The j-th level value of C (denoted by $\text{level}_j(C)$) corresponds to the (unique) node on level j of the ascending branch starting from C. If j is higher than the level of C then $\text{level}_j(C)$ is considered undefined. By assigning to each level a weight we can obtain a similarity measure, s_α, which generalizes the Wu-Palmer similarity. This measure is described in Eq.(7) where $\alpha_i \in [0,1]$ such that $\alpha_1 + \alpha_2 + \ldots \alpha_k = 1$ and $\sigma_j(C_1, C_2)$ is defined in Eq.(8). As s_α takes values in $[0,1]$ the corresponding dissimilarity is $\delta_\alpha = 1 - s_\alpha$.

C_1	C_2	$lca(C1, C2)$	$s_{WP}(C_1, C_2)$	$s_{Lin}(C_1, C_2)$	$s_{SBI}(C_1, C_2)$	$s_\alpha(C_1, C_2)$	
A00.0	A00.0	A00.0	1	1	1	$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$	same code
A00.0	A00.9	A00	0.75	0.84	0.94	$\alpha_1 + \alpha_2 + \alpha_3$	same section
A00.0	A09.0	A00-A09	0.5	0.76	0.67	$\alpha_1 + \alpha_2$	same group
A00.0	B99	ch.I	0.25	0.64	0.35	α_1	same chapter
A00.0	Z00.8	ICD-10	0	0	0	0	different chapters

Table 1: Similarity values for two entities from the ICD-10 taxonomy (Figure 1). The probabilities involved in the computation of s_{Lin} are: $P(A00.0) = 0.00005$, $P(A00.9) = 0.00009$, $P(A090) = 0.00014$, $P(B99) = 0.00008$, $P(A00 - A09) = 0.00080$, $P(A00) = 0.00030$, $P(ch.I) = 0.00200$. The number of leaves counted on the ICD-10 tree which are used in the computation of s_{SBI} are: $|\textbf{leaves}(A00)| = 3$, $|\textbf{leaves}(A00 - A09)| = 59$, $|\textbf{leaves}(ch.I)| = 777$, $|\textbf{tree leaves}| = 9573$.

$$s_\alpha(C_1, C_2) = \sum_{i=1}^{k} \alpha_i \prod_{j=1}^{i} \sigma_j(C_1, C_2) \tag{7}$$

$$\sigma_j(C_1, C_2) = \begin{cases} 1 & \text{if level}_j(C_1) = \text{level}_j(C_2) \\ 0 & \text{if level}_j(C_1) \neq \text{level}_j(C_2) \\ & \text{or at least one level is undefined} \end{cases} \tag{8}$$

In the case of ICD-10 codes, the first level is represented by the chapters, the second level is represented by the groups, the third level corresponds to sections and the fourth to codes (under the assumption that only four levels are taken into account). Examples of similarity values computed using s_α are presented in Table 1. By choosing appropriately the values of parameters α_i one can control the influence of each level.

PROPOSITION 1. *The similarity measure, s_α, and the corresponding dissimilarity measure, δ_α, satisfy the following properties:*

(i) for $\alpha_i = 1/k$, s_α is identical to the Wu-Palmer similarity for leaf entities corresponding to the same level;

(ii) d_α satisfies the strong triangle inequality, i.e. $d_\alpha(C_1, C_2) \leq \max\{d_\alpha(C_1, C_3), d_\alpha(C_2, C_3)\}$ for any triple (C_1, C_2, C_3);

(iii) if $\alpha_{i^*} = 1$ then s_α is a binary measure which takes into account only the levels up to i^*; if $i^* = k$ then the simple binary similarity is obtained (which is 1 only when the full entities are identical and is 0 in all the other cases).

PROOF. (i) If C_1 and C_2 are two entities placed on level k of the taxonomy and they are identical up to level i then $s_\alpha = \sum_{j=1}^{i} \alpha_j$. If $\alpha_j = 1/k$ for $j = \overline{1, k}$ then $s_\alpha(C_1, C_2) = i/k$. On the other hand, if C_1 and C_2 are identical up to level i then $\text{depth}(lca(C_1, C_2)) = i$, hence $s_{WP}(C_1, C_2) = i/k$. Thus the Wu-Palmer similarity is a particular case of the weighted similarity characterized by equal weights.
(ii) Let us suppose that there exists at least a triple (C_1, C_2, C_3) such that $\delta_\alpha(C_1, C_2) > \delta_\alpha(C_1, C_3)$ and $\delta_\alpha(C_1, C_2) > \delta_\alpha(C_2, C_3)$. This would mean that

$$\sum_{i=1}^{k} \alpha_i \prod_{j=1}^{i} \sigma_j(C_1, C_2) < \sum_{i=1}^{k} \alpha_i \prod_{j=1}^{i} \sigma_j(C_1, C_3)$$

and

$$\sum_{i=1}^{k} \alpha_i \prod_{j=1}^{i} \sigma_j(C_1, C_2) < \sum_{i=1}^{k} \alpha_i \prod_{j=1}^{i} \sigma_j(C_2, C_3).$$

In order to have these inequalities satisfied it would be necessary to exist $i' \leq k$ and $i'' \leq k$ such that

(a) $\sigma_{i'}(C_1, C_2) = 0$ and $\sigma_{i''}(C_1, C_2) = 0$;

(b) $\sigma_j(C_1, C_3) = 1$ for any $j \leq i'$ and $\sigma_j(C_2, C_3) = 1$ for any $j \leq i''$.

If $i' = i''$, condition (a) implies that $\text{level}_{i'}(C_1) \neq \text{level}_{i'}(C_2)$ and condition (b) implies that $\text{level}_{i'}(C_1) = \text{level}_{i'}(C_3)$ and $\text{level}_{i'}(C_2) = \text{level}_{i'}(C_3)$ which leads to a contradiction. If $i' \neq i''$, let us suppose that $i' < i''$. Thus $\text{level}_{i'}(C_1) \neq \text{level}_{i'}(C_2)$ and $\text{level}_{i'}(C_1) = \text{level}_{i'}(C_3)$ and $\text{level}_{i'}(C_2) = \text{level}_{i'}(C_3)$ which leads again to a contradiction. A similar result is obtained if $i' > i''$. Thus, for any triple (C_1, C_2, C_3) the strong triangle inequality is satisfied.
(iii) If $\alpha_{i^*} = 1$ then it means that all other weights are zero. Thus $s_\alpha(C_1, C_2) = 1$ if $level_j(C_1) = level_j(C_2)$ for each $j \leq i^*$ and $s_\alpha(C_1, C_2) = 0$ in all the other cases. □

4.2 Role of weights

The weighted similarity allows to generalize the edge- counting measures by allowing different edges to participate differently in the computation of the similarity. In this way different importance can be assigned to different levels in a hierarchical taxonomy. The extreme cases when only one weight (α_{i^*}) is non-zero correspond to "compressed" taxonomies in which all levels higher than i^* are "absorbed" in the i^*-th level. One of the main questions is which are the appropriate weights to use for a given data analysis task. This problem is similar to that of feature selection/weighting for which a lot of studies had been already conducted. Following the line of reasoning from the feature selection/weighting problem the weights search problem can be formulated as an optimization of some quality criteria which can be estimated using some information on the ground truth structure in the data. It should be mentioned that sometimes we are interested in avoiding binary values for the weights (since this would generate binary similarity values). In such cases the natural approach is to combine the quality criteria with a regularization term aiming to penalize weight values approaching 1 or 0. In the experimental analysis presented in Section 5 we analyzed both the case when the weights can take any value in $[0, 1]$ and the case when extreme values are penalized.

The main advantage of the weighted similarity is that, once the weights are established, the similarity is easy to evaluate and depends only on the level-structure of the taxonomy, being robust with respect to changes in the taxonomy which do not alter the level structure (unlike the measures based on the number of leaves which could be sensitive to small changes in the taxonomy). On the other hand, this

simplicity of the weighted similarity may limit its discriminant capacity since it does not take into account estimations of generality or concreteness of involved concepts. The effectiveness of the proposed measure in estimating the similarities between lists of diagnostic codes is analyzed for medical data sets in section 5.

5. ASSESSING THE GOODNESS OF A SIMILARITY MEASURE

The quality of a similarity measure can be evaluated either by direct or by indirect methods [13]. The direct methods are based on analyzing the correlation between the similarity estimated using the analyzed measure and some previously known similarity values. This approach is typically used in the evaluation of semantic similarity between concepts using similarity scores provided by human experts (see for instance [8]). The indirect methods evaluate the quality of a similarity by assessing the results of a classification or clustering task which involves the analyzed similarity measure. In this last case the assessment is based on quality measures specific to the task used as support of the analysis (e.g. classification accuracy indicators or goodness of clustering indicators). The results provided by indirect methods are influenced by the performance of the algorithm used to solve the classification or clustering task, thus they may not reflect accurately the quality of the similarity measure, independently of a specific task solving method.

In this paper we use the first approach but since we do not know similarity scores for specific pairs of elements (e.g. lists of diagnostic codes corresponding to pairs of patients) we shall consider as ground truth partitions of data which incorporate intrinsically the knowledge on the similarity provided by the experts. Thus instead of analyzing the correlation between the estimated similarity scores and those provided by the human experts we will assess how well can explain the analyzed similarity measure the existing data partition. Therefore we shall evaluate the similarity by using internal cluster validity indices, which analyze the compactness of clusters and their separation.

The clustering validity indicators are typically used to assess the performance of a clustering method or to identify the proper number of clusters in a context when the similarity measure is known. Here we use them in a different context: we evaluate a similarity measure by analyzing how good would be the observed data partition if it would be obtained using the analyzed similarity. Thus, good clustering quality would suggest appropriate similarity measure. The problem which arises now is to choose a clustering quality indicator. The choice we made is based on the results of the evaluation methodology proposed in [5] which suggests that C-Index obtained the highest score when compared with 6 other indices in an experimental study involving 11 datasets.

For a data partition $\mathcal{P} = \{P_1, P_2, ..., P_K\}$ containing K clusters, C-Index is computed by following the steps:
(i) Compute the sum V of dissimilarities between all elements belonging to the same cluster, i.e. $V = \sum_{k=1}^{K} V_k$, $V_k = \sum_{x,y \in P_k} d(x, y)$.
(ii) Compute the dissimilarities between all distinct elements in the data set and sort increasingly the list of dissimilarities (L).
(iii) Compute V_{min} as the sum of the first r elements in the list L and V_{max} as the sum of the last r elements of L (r represents the number of distances corresponding to the pairs belonging to same clusters and which were involved in the computation of V).
(iv) The value of C-Index is $(V - V_{min})/(V_{max} - V_{min})$.

C-Index takes values in $[0, 1]$ and as its value is smaller, the clustering quality is higher.

6. EXPERIMENTAL ANALYSIS

The practical context of this study is related to the problem of identifying risk factors for preterm birth using information from medical records of mothers and their corresponding newborn. Therefore the analysis is conducted based on two datasets collected from obstetrics and neonatology wards of two Romanian hospitals.

6.1 Data sets

The source data collected through the Diagnosis Related Groups (DRG) system were first pre-processed by following the framework described in [2] in order to aggregate information of mothers and corresponding newborns. This aggregation resulted in 3176 records in the first dataset and 1655 records in the second dataset containing lists of diagnostic and medical procedures codes for (mother, newborn) pairs. In a next step, each data set has been partitioned in groups based on the main diagnostic or medical procedure applied to the newborn. Consequently, in each data set has been identified a group containing information on mothers whose newborn does not have any pathology (normal group) and several groups corresponding to newborns with various abnormal health conditions (e.g. "disorders related to length of gestation and fetal growth", "haematological disorders", "congenital malformations" etc.). Fourteen such groups have been identified in the first dataset and sixteen in the second one. The synthetical description of the partitions used as ground truth in evaluating the similarity measures, obtained after eliminating the ambiguous cases (same list of diagnostic codes in the mother record and different pathological conditions in the newborn record) is provided in Table 2.

6.2 Research questions and methodology

The experimental study has been designed in order to address the following questions:

Q1. Is the proposed weighted similarity measure competitive with respect to edge-counting and information-content based measures? Do the values of weights have a significant influence on the results?

Q2. Is the discriminant ability of a dissimilarity significantly influenced by the set-level dissimilarity (e.g. standard Hausdorff or Dice variant)?

In order to get an answer to the first question we formulated the problem of weights estimation as an optimization one. The search space is represented by $[0, 1]^k$ and the objective function is the value of a cluster validity index computed using δ_α as code-level dissimilarity and the observed partition as ground truth. To solve this optimization problem we used a recent variant of SHADE (Success History Adaptive Differential Evolution) [10], one of the most competitive evolutionary algorithms. The analysis has been conducted using both datasets presented in the previous subsection. Both the full data partition (including the normal group and

Characteristic	Data set 1				Data set 2			
	all data group	normal group	preterm group	low weight group	all data group	normal group	preterm group	low weight group
Number of items	1343	374	85	56	859	401	46	33
Number of distinct codes	229	64	65	52	160	56	49	36
List length	4.9±1.4	4.4±1.3	5.3±1.2	5.2±1.3	4.3±1.3	4.1±1.2	4.7±1.4	4.3±1.1

Table 2: Synthetic characteristics of the datasets containing lists of ICD-10 diagnostics and medical procedures codes. The number of items (an item is a list of codes corresponding to a patient), the number of distinct ICD-10 codes and the average and standard deviation of the number of codes per list are reported for the entire set of items (corresponding to health records of mothers who gave birth to children) and for three groups of cases: normal cases, cases of pre-term births (section P07 in ICD-10 taxonomy) and cases of children with low birth weight (section P05 in ICD-10 taxonomy).

the other groups corresponding to various newborn pathologies) and two particular cases (involving the normal group and groups of preterm birth and low weight cases) have been used to assess the similarity quality.

As code-level (dis)similarity measures, besides the proposed weighted measure, other three measures have been involved in the study: (i) Wu-Palmer measure (δ_{WP}), which is a particular case of the weighted one; (ii) Lin measure (δ_{Lin}), which uses probabilities estimated based on the data corpus represented by the joined datasets; (iii) Sanchez-Batet-Isern measure (δ_{SBI}) which is based on estimating the information content as described in Eq. (4) using the content of the ICD-10 taxonomy.

In order to address the second question each of the four code-level dissimilarities has been combined with two set-level dissimilarities: the Hausdorff distance (Eq.5) and a variant of this distance, which in discrete spaces is similar to the Dice dissimilarity (Eq.6). Thus the comparative analysis involved eight dissimilarity measures between lists of ICD-10 codes.

The evaluation framework was implemented in Java and the Java version of SHADE (with default values of the parameters) provided by its authors[5] has been used.

6.3 Results and discussion

A first set of results corresponding to three types of observed data partitions are presented in Table 3 where values of the C-Index are provided. For each data partition the bolded values correspond to the dissimilarity with smallest C-Index value (out of all combinations of set-level and code-level measures) and the italic ones correspond to the best code-level dissimilarity for each group of measure. In the case of the weighted dissimilarity (δ_α) are provided results for all cases of binary weights (which lead to binary (dis)similarity values), for the weight vector (α^*) estimated using the evolutionary algorithm SHADE in the case when no constraint is imposed on the weights values and for the weight vector (α_r) obtained in the case when a regularization term was used in the criteria to be minimized (e.g. $(CIndex(\delta_\alpha) - \prod_{i=1}^{k}(1-\alpha_i))/2$). It should be noticed that in the unconstrained case the estimated optimal weights are in $(0,1)$ but in many cases they are very close to the border of the search space (e.g. $(1.65\cdot10^{-6}, 2.77\cdot10^{-5}, 0.9996, 3.59\cdot10^{-4})$). However such values are reported as binary ones in Table 3, e.g. $\alpha^* \simeq (0,0,1,0)$. On the other hand, in several cases the evolutionary approach does not provide the best value in the allocated computational budget (2000 generations). The main remarks which can be inferred from the results presented in Table 3 are:

(i) In almost all cases the smallest C-Index value is obtained by the weighted similarity with binary weights (or close to binary). For most datasets the discriminant level (that for which the weight is highest) is the third (section) or fourth (code) level. In the case when the binary weights are penalized, the C-index value is significantly higher.

(ii)In the case of two-groups partitions (normal vs. preterm birth and normal vs. low weight) the combination between the Hausdorff distance (d_H) and the optimized weighted dissimilarity leads to C-Index values equal to 0 because the set of dissimilarity values is $\{0,1\}$ and the group of normal cases contains several subgroups of identical lists of diagnostic codes (groups of mothers with identical lists of diagnostics and procedures).

(iii) When the Dice variant of the set-level measure is used (d_D) the binary weights are no more leading always to the best results (especially for the two groups partitions). However the weighted dissimilarity still leads to the smallest C-index values.

(iv) In the case of full datasets, the dissimilarity based on the taxonomy structure (δ_{SBI}) leads to better results than the other two (δ_{WP} and δ_{Lin})

Based on these remarks one can say that the answers to questions Q1 and Q2 are both affirmative, as the weighted dissimilarity obtained better quality values for all datasets but its performance is highly influenced by the values of weights and the set-level dissimilarity. By tuning the weights to a dataset one can see that there exist dissimilarity measures in agreement with the existing partition, but such tuned measures should be evaluated with respect to their generalization ability before using them in practice. The relative superiority of the dissimilarity with (almost) binary weights should be interpreted with caution, as C-Index could be biased in the case of binary measures (as those obtained by combining d_H with binary code-level dissimilarities). In order to analyze the influence of the cluster validity index on the dissimilarity measures assessment we conducted an analysis for two other well-known indices: Silhouette (to be maximized) and Davies-Bouldin (to be minimized). The results obtained for the partition consisting of normal birth and preterm birth groups are presented in Table 4. The main remarks inferred from these results are:

[5]https://sites.google.com/site/tanaberyoji/home

Set level	Code level	C-Index for data set 1			C-Index for data set 2		
		all groups	normal vs. preterm	normal vs. low weight	all groups	normal vs. preterm	normal vs. low weight
d_H	δ_{1000}	0.373	0.546	0.559	0.412	0.512	0.609
	δ_{0100}	0.862	0.233	0.201	0.690	0.134	0.141
	δ_{0010}	0.582	0.075	0.033	0.338	0.027	0.044
	δ_{0001}	*0.284*	**0.0**	**0.0**	**0.060**	**0.0**	**0.0**
	δ_{α^*}	0.364	0.038	0.026	0.118	0.014	0.023
	$\alpha^* \simeq$	(0.04,0,0.01,0.95)	(0,0,0.12,0.88)	(0,0,0.43,0.57)	(0,0,0.04,0.96)	(0,0,0.2,0.8)	(0,0.02,0.18,0.8)
	δ_{α_r}	0.547	0.406	0.389	0.590	0.385	0.390
	$\alpha_r \simeq$	(0.42,0, 0.26,0.32)	(0,0, 0.5,0.5)	(0,0,) 0.5,0.5)	(0,0,) 0.42,0.58)	(0,0.01, 0.48,0.51)	(0,0.03, 0.47,0.5)
	δ_{WP}	0.576	0.446	0.415	0.474	0.358	0.385
	δ_{Lin}	0.574	*0.423*	*0.401*	0.476	*0.321*	*0.351*
	δ_{SBI}	*0.500*	0.480	0.465	*0.448*	0.413	0.458
d_D	δ_{1000}	0.186	0.490	0.598	0.234	0.560	0.695
	δ_{0100}	0.365	0.366	0.429	0.331	0.352	0.426
	δ_{0010}	0.374	0.359	0.426	0.375	0.348	0.435
	δ_{0001}	0.458	0.327	0.393	0.399	0.300	0.339
	δ_{α^*}	**0.186**	*0.326*	*0.391*	*0.234*	*0.295*	*0.339*
	$\alpha^* \simeq$	(1,0,0,0)	(0,0.24,0,0.76)	(0,0.17,0.05,0.78)	(1,0,0,0)	(0,0.19,0,0.81)	(0,0.04,0,0.96)
	δ_{α_r}	0.517	0.512	0.545	0.507	0.501	0.536
	$\alpha_r \simeq$	(0.47,0.26, 0.18,0.09)	(0.20,0.25, 0.24,0.31)	(0.18,0.25, 0.26,0.31)	(0.36,0.30, 0.18,0.16)	(0.19,0.27, 0.21,0.33)	(0.15,0.27, 0.18,0.40)
	δ_{WP}	0.371	*0.346*	*0.412*	0.338	*0.330*	0.408
	δ_{Lin}	0.413	0.348	0.417	0.358	0.331	*0.398*
	δ_{SBI}	*0.327*	0.362	0.433	*0.316*	0.350	0.447

Table 3: Values of C-Index corresponding to various combinations of the ICD-10 taxonomy based dissimilarity with set-based dissimilarities. The evaluation is based on the overall partitions (14 groups in data set 1 and 16 groups in data set 2) and partitions containing two groups (normal vs. pre-term birth cases and normal vs. low weight at birth cases). α^* denotes unconstrained (sub)optimal values of the weights while α_r denotes regularized (sub)optimal weights' values.

Set level	Code level	Data set 1 (normal vs preterm)			Data set 2 (normal vs preterm)		
		C-Index (↓)	Silhouette (↑)	Davies-Bouldin (↓)	C-Index (↓)	Silhouette (↑)	Davies-Bouldin (↓)
d_H	δ_{α^*}	**0.038**	0.043	1.859	**0.014**	0.037	1.927
	$\alpha^* \simeq$	(0,0,0.12,0.88)	(0.43,0,0.57,0)	(1,0,0,0)	(0,0,0.2,0.8)	(0.02,0,0.98,0)	(0.01,0,0.99,0)
	δ_{α_r}	0.406	0.035	1.917	0.385	0.028	1.938
	$\alpha_r \simeq$	(0,0, 0.5,0.5)	(0.26,0.22, 0.31,0.21)	(0.36,0.17, 0.28,0.13)	(0,0.01, 0.48,0.51)	(0.26,0.25, 0.27,0.22)	(0.29,0.24, 0.34,0.13)
	δ_{WP}	0.446	0.035	1.924	0.358	0.028	1.942
	δ_{Lin}	*0.423*	0.023	1.951	*0.321*	0.025	1.949
	δ_{SBI}	0.480	*0.042*	*1.904*	0.413	*0.031*	*1.933*
d_D	δ_{α^*}	*0.326*	**0.139**	**1.510**	*0.295*	**0.114**	**1.616**
	$\alpha^* \simeq$	(0,0.24,0,0.76)	(1,0,0,0)	(0.98,0.02,0,0)	(0,0.19,0,0.81)	(1,0,0,0)	(0.98,0.02,0,0)
	δ_{α_r}	0.512	0.092	1.796	0.501	0.075	1.844
	$\alpha_r \simeq$	(0.20,0.25, 0.24,0.31)	(0.27,0.31, 0.27,0.15)	(0.35,0.25, 0.21,0.19)	(0.19,0.27, 0.21,0.33)	(0.30,0.22, 0.27,0.21)	(0.31,0.26, 0.24,0.19)
	δ_{WP}	*0.346*	0.089	1.811	*0.330*	0.073	1.853
	δ_{Lin}	0.348	0.076	1.842	0.331	0.066	1.870
	δ_{SBI}	0.362	*0.103*	*1.764*	0.350	*0.083*	*1.818*

Table 4: Values of C-Index, Silhouette and Davies-Bouldin indicators corresponding to various combinations of the ICD-10 taxonomy based dissimilarity with set-based dissimilarities. Smaller values for C-Index and Davies-Bouldin indices and larger values for the Silhouette index suggest better agreement between the (dis)similarity measure and the observed partition.

(i) The best validity indices are still obtained for the weighted measure but in the case of Silhouette and Davies-Bouldin indices the usage of Dice variant produces better results than the standard Hausdorff distance.

(ii) When the weighted dissimilarity is combined with d_D the most discriminant level seems to be the first one (the chapter level), which would mean that in order to discriminate preterm cases it would be enough to use only the chapter code (the disease family).

(iii) With respect to Silhouette and Davies-Bouldin indices, the dissimilarity based on the taxonomy structure (δ_{SBI}) leads to better results than Wu-Palmer (δ_{WP}) and Lin (δ_{Lin}) dissimilarities, thus it could be used as an alternative to the weighted dissimilarity.

The relative superiority of binary measures observed in the experiments was not expected and finding an explanation of this is an open question.

7. CONCLUSIONS AND FURTHER WORK

Aiming to identify appropriate (dis)similarity measures between patient profiles containing lists of diagnostic codes we analyzed several taxonomy-based measures and proposed a weighted version of the edge-counting similarity. The proposed code-level measure satisfies the strong triangle inequality and generalizes both the simple binary measure as well as the Wu-Palmer distance. In order to extend the code-level dissimilarity to lists of codes we combined it with two variants of the Hausdorff distance between sets. Aiming to assess the quality of a dissimilarity measure we proposed to estimate how well the measure can explain observed data partitions using known cluster validity indices. Using an evolutionary algorithm we estimated the weight values which optimizes the score provided by a cluster quality index.

The experimental analysis conducted for datasets collected from obstetrics and neonatology wards of two hospitals revealed the fact that the weighted code-level (dis)similarity leads to better scores than other taxonomy-based measures. However this measure is sensitive with respect to the weights values and the obtained results are influenced by the used cluster validity index. The dissimilarity based on the taxonomy structure (δ_{SBI}) could be a practical alternative, as it does not require parameter tuning. On the other hand, there is not enough evidence to prove that one of the set-level distances is better than the other one.

Assessing the quality of the dissimilarity measure in a multi-criterial framework by using several validity indices (e.g. C-Index, Silhoutte, Davies-Bouldin etc.) is one of the lines of further research. On the other hand, further extensions involving hybridization of different taxonomy-based measures and their evaluation in the context of full patient profiles (containing other medical information besides the list of diagnostics) are to be addressed in future.

8. REFERENCES

[1] L. Ceccaroni and L. Subirats. Interoperable knowledge representation in clinical decision support systems for rehabilitation. *Appl. Comput. Math.*, Il (Special issue)(2):303–316, 2012.

[2] R. Dogaru, D. Zaharie, D. Lungeanu, E. Bernad, and M. Bari. A framework for mining association rules in data on perinatal care. In *The 8th International Conference on Technical Informatics*, CONTI '08, pages 147–152, 2008.

[3] 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.

[4] M. Golfarelli and E. Turricchia. A characterization of hierarchical computable distance functions for data warehouse systems. *Decision Support Systems*, 62:144–157, 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] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, 1998.

[8] 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.

[9] 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.

[10] T. Ryoji and F. Alex. Improving the search performance of shade using linear population size reduction. In *2014 IEEE Congress on Evolutionary Computation (CEC)*, CEC'14, pages 1658–1665, 2014.

[11] D. Sánchez 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] G. Srinivas, N. Tandon, and V. Varma. A weighted tag similarity measure based on a collaborative weight model. In *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pages 79–86, 2010.

[14] L. Subirats, L. Ceccaroni, and F. Miralles. Knowledge representation for prognosis of health status in rehabilitation. *Future Internet*, 4(2):762–775, 2012.

[15] J. Sun, F. Wang, J. Hu, and S. Edabollahi. Supervised patient similarity measure of heterogeneous patient records. *ACM SIGKDD Explorations Newsletter*, 14(1):16–24, 2012.

[16] 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.

[17] W. T. Yih. Learning term-weighting functions for similarity measures. In *Proceedings of EMNLP: Volume 2*, pages 793–802, 2009.