Journal of Biomedical Informatics 63 (2016) 66–73

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Using concept hierarchies to improve calculation of patient similarity

Dominic Girardi [a], Sandra Wartner [a], Gerhard Halmerbauer [b], Margit Ehrenmüller [b], Hilda Kosorus [c], Stephan Dreiseitl [d,*]

[a] RISC Software GmbH, Research Unit Medical Informatics, Johannes Kepler University Linz/Hagenberg, Austria
[b] Department of Process Management in Health Care, University of Applied Sciences Upper Austria, Steyr, Austria
[c] Institute for Application Oriented Knowledge Processing, Johannes Kepler University Linz, Austria
[d] Department of Software Engineering, University of Applied Sciences Upper Austria, Hagenberg, Austria

ARTICLE INFO

Article history:
Received 4 March 2016
Revised 31 May 2016
Accepted 26 July 2016
Available online 28 July 2016

Keywords:
Distance measure using concept hierarchy
ICD-10 taxonomy
Patient similarity calculation

ABSTRACT

Objective: We introduce a new distance measure that is better suited than traditional methods at detecting similarities in patient records by referring to a concept hierarchy.
Materials and methods: The new distance measure improves on distance measures for categorical values by taking the path distance between concepts in a hierarchy into account. We evaluate and compare the new measure on a data set of 836 patients.
Results: The new measure shows marked improvements over the standard measures, both qualitatively and quantitatively. Using the new measure for clustering patient data reveals structure that is otherwise not visible. Statistical comparisons of distances within patient groups with similar diagnoses shows that the new measure is significantly better at detecting these similarities than the standard measures.
Conclusion: The new distance measure is an improvement over the current standard whenever a hierarchical arrangement of categorical values is available.

1. Introduction

The ubiquitous availability of data processing systems has led to an ever-increasing amount of data in healthcare environments [1,2]. While some of this data is available only in unstructured formats and thus difficult to process automatically—recent developments in image understanding [3] and natural language processing [4] notwithstanding—several areas of biomedicine have seen great strides towards standardized formats for data representation. Examples of such formats include DICOM for imaging data, Gene Ontology for molecular biology terms, SNOMED CT for clinical terminology, and ICD-10 for the classification of diseases.

While standardized data formats were mainly developed to facilitate information exchange between computers, structured data representation also provides benefits to the human understanding of data. These benefits may be minor—for example, providing descriptive summary statistics of the data—or major, when visualizing complex gene activation pathways in cells. The ability to graphically represent data may in fact be the biggest advantage of structured data formats, because it allows the human pattern recognition apparatus to derive meaning from the data [5,6]. In many instances, this requires a similarity measure to be defined on the data, so that similarities in the original data space can be mapped (in a meaningful manner) to a 2D representation on screen.

One context in which similarity information can be obtained from data is when the data concepts are arranged in a hierarchical manner [7,8]. An example of such a concept hierarchy is the International Classification of Diseases catalog ICD-10, maintained by the World Health Organization [9]. It contains over 12,000 disease classifications organized in three levels, with 22 level 1 elements.

Consider how a concept hierarchy can help to detect similarities between two patients A and B. Patient A suffers from *influenza due to identified avian influenza virus* (ICD-10 code J09.0) and a *fracture of fibula alone* (ICD-10 code S82.4). Patient B suffers from *pneumonia* (ICD-10 code J18.9) and a *fracture of lateral malleolus* (ICD-10 code S82.6). Regarding these diagnoses only as nominal (categorical) values, patient A shows no similarity to patient B, because their diseases and disease codes are all different. From a real-world point of view, however, it is obvious that the diagnoses of both patients are quite similar, because influenza is similar to pneumonia, and a broken fibula is similar to a broken malleolus. What is needed to accurately reflect this similarity is a distance measure that takes into account the hierarchical structure of the

* Corresponding author.
E-mail address: stephan.dreiseitl@fh-hagenberg.at (S. Dreiseitl).

http://dx.doi.org/10.1016/j.jbi.2016.07.021
1532-0464/

concept hierarchy. Furthermore, this measure must also calculate the distance between *sets* of concepts.

The work presented here is motivated by the problem of finding and visualizing similarities in categorical clinical patient data, where every patient is represented as a set of ICD-10 codes. Two patients are considered as similar if they show similar or overlapping sets of diagnoses. The main hypothesis is that patients with similar diagnoses (meaning the same diagnoses on a high level of the hierarchy) form clearer clusters when the hierarchical structure is incorporated into the distance measure than when it is not.

We will provide two ways to show that this hypothesis holds. First, we provide a graphical representation of clusterings and show that patients with similar diagnoses form clearly visible clusters with the new hierarchical distance measure, while they do not when the hierarchical structure is ignored. Second, we calculate the inter-record distances of patients with similar diagnoses (same diagnoses on ICD-10 level 2) and compare the results of the new hierarchical distance measure with a standard distance measures.

The improved, more realistic distance calculation contributes to a number of applications. The data which is presented in this paper is taken from a clinical benchmarking program. The improved distance calculation allows a more accurate visualization of the benchmarking data and subsequently more reliable and understandable benchmarks. Generally, distance-based data visualization methods (e.g. dimensionality reduction or non-linear mappings) are important tools for exploratory data analysis. A more accurate distance measure leads to higher expressiveness of the resulting data visualizations.

Moreover, the calculation of distance or similarity between two data sets is at the core of any case-based reasoning approach [10], with many applications in biomedicine – one example being decision support systems. Since such systems depend so heavily on reasoning by similarity (using similar old cases to reason about new ones), improvements in the assessment of case similarities directly lead to improvements in the capabilities of such systems.

2. Related work

The notions of *patient distance* and *patient similarity* have been widely investigated in the biomedical literature. In this work, we focus on the notion of *semantic similarity*, where the semantics of concepts are arranged in a hierarchical manner. Two problems arise in this context: How to measure the distance between individual concepts, and how to measure the distance between sets of concepts.

2.1. Semantic similarity between concepts

One can distinguish two ways of using an ontology or taxonomy to determine the semantic similarity between concepts: the *edge-based approach* and the *information content-based approach* [11]. Our approach, however, deals only with hierarchical structure in the data, and ignores frequency distributions in the data set.

The similarity, dissimilarity or distance between two concepts in a hierarchy is calculated by considering the hierarchy tree as an acyclic graph and applying graph distance measures to it. Boutsinas and Papastergiou [12] present an algorithm that calculates the distance between two concepts in an hierarchy by the hierarchy level of their nearest common ancestor node. The lower the nearest common ancestor node is located in the tree (with the tree root considered as the top), the more similar are the two concepts. A similar approach can be observed in the work of Hammer et al. [13], who extend the concept of a Self-Organizing Map (SOM) [14]. Their generalized SOM treats any enumeration as a hierarchy; flat lists are hierarchies with only one level.

Intuitively, the similarity of different concepts in an ontology is measured by computing their edge distance within the ontology. This means that the closer two concepts are in the ontology, the more similar we consider them to be [15]:

$$\mathrm{sim}(c_1, c_2) = \text{minimum number of edges separating } c_1 \text{ and } c_2,$$

where c_1 and c_2 are the node representation of the two concepts in the ontology. Wu and Palmer [16] redefined the edge-based similarity measure taking into account the depth of the nodes in the hierarchical graph:

$$\mathrm{sim}(c_1, c_2) = \frac{2N_3}{N_1 + N_2 + 2N_3}, \tag{1}$$

where N_1 and N_2 are the number of nodes from c_1 and c_2, respectively, to c_3, the *least common superconcept* (LCS) of c_1 and c_2, and N_3 is the number of nodes on the path from c_3 to the root node.

Li et al. [17] defined the similarity between two concepts as:

$$\mathrm{sim}(c_1, c_2) = \begin{cases} e^{-\alpha(N_1+N_2)} \cdot \frac{e^{\beta N_3} - e^{-\beta N_3}}{e^{\beta N_3} + e^{-\beta N_3}} & \text{if } c_1 \neq c_2, \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

where the parameters α and β scale the contribution of the two values $N_1 + N_2$ and N_3. On a benchmark data set, they obtained the optimal parameters settings as $\alpha = 0.2$ and $\beta = 0.6$.

The *information content*-based approach for computing the semantic similarity between concepts was introduced by Resnik [18]. It assumes that the frequency with which one term appears with another within a given ontology represents the similarity of the two terms. Resnik [19] showed that by associating probabilities with concepts in the taxonomy, it is possible to capture the same idea of edge-based similarity, but avoid the unreliability of uniform edges.

Resnik [18] defines the similarity of two concepts as

$$\mathrm{sim}(c_1, c_2) = \max_{c_3 \in S(c_1, c_2)} -\log(p(c_3)),$$

where $S(c_1, c_2)$ is the sets of all superconcepts of c_1 and c_2, and $p(c_3)$ is the relative frequency of concept c_3.

Compared to the edge-counting method, the similarity measure introduced by Resnik [19] is conceptually quite simple. However, it is not sensitive to the problem of varying link distances. In addition, by combining an ontological structure with empirical probability estimates, it provides a way of adapting a static knowledge structure to multiple contexts.

This similarity measure was further improved by Lin [20], when he introduced the information-theoretic definition of similarity. Based on this notion, he defined the semantic similarity in a taxonomy as

$$\mathrm{sim}(c_1, c_2) = \frac{2 \times \log(p(c_3))}{\log(p(c_1)) + \log(p(c_2))},$$

where c_3 is the LCS of c_1 and c_2. Here, one can notice the similarities with the measure in Eq. (1).

2.2. Semantic similarity between sets of concepts

Defining a semantic similarity measure between sets of concepts was the next step in computing semantic similarity mainly for information retrieval purposes.

In Bouquet et al. [21], the ontological distance between sets of concepts X and Y is computed by summing up the distances between every pair (c_1, c_2), where $c_1 \in X$ and $c_2 \in Y$. Haase et al. [22] used the edge-based similarity measure between concepts defined in Eq. (2) to introduce the similarity between sets of concepts as