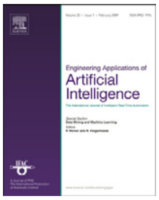
Engineering Applications of Artificial Intelligence 24 (2011) 1398–1409

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



A SNOMED supported ontological vector model for subclinical disorder detection using EHR similarity

L.W.C. Chan [a],*, Y. Liu [b], C.R. Shyu [c], I.F.F. Benzie [a]

[a] *Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong*
[b] *Department of Mechanical Engineering, National University of Singapore, Singapore 117576, Singapore*
[c] *MU Informatics Institute, The University of Missouri, USA*

ARTICLE INFO

Available online 15 June 2011

Keywords:
Electronic Health Record
Vector model
SNOMED
Similarity
Clinical decision support
Atherosclerosis

ABSTRACT

Electronic Health Records (EHR) form a valuable resource in the healthcare enterprise because clinical evidence can be provided to identify potential complications and support decisions on early intervention. Simple string matching, the common search algorithm, is not able to map a query to the similar health records in the database with respect to the medical concepts. A novel ontological vector model supported by the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) is proposed in this paper to project the disease terms of a health record to a feature space so that each health record can be characterized using a feature vector, giving a fingerprint of the record. The similarity between the query and database health records was measured by similarity measures of their feature vectors and string matching score respectively. Three types of similarity measures were considered in this study, namely, Euclidean distance (ED), direction cosine (DC) and modified direction cosine (mDC). Medical history and carotid ultrasonic imaging findings were collected from 47 subjects in Hong Kong. The dataset formed 1081 pairs of health records and ROC analysis was used to evaluate and compare the accuracy of the ontological vector model and simple string matching against the agreement of the presence or absence of carotid plaques identified by carotid ultrasound between two subjects. It was found that the score generated by simple string matching was a random rater but the ontological vector model was not. In other words, the degree of health record similarity based on the ontological vector model is associated with the agreement of atherosclerosis between two patients. The vector model using feature terms at the SNOMED-CT level 4 gave the best performance. The performance of mDC was very close to that of ED and DC but the properties of mDC make it more suitable for the retrieval of similar health records. It was also shown that the ontological vector model was enhanced by the support vector classifier approach.

1. Introduction

Electronic Health Record (EHR) system is patient-centered information resource supported by computer software and hardware infrastructure. In the clinical workflow, EHR system facilitates the archiving and communications of clinical information of each patient throughout the episodes of care. More than a data repository of digitized health records, EHR system provides core functionalities in eight categories, including electronic health information/data capture, order entry/order management, results management, administrative processes, electronic connectivity, clinical decision support (CDS), health outcomes reporting and patient support (Institute of Medicine and Board on Health Care Services, 2003). Among these eight categories, clinical decision support enhances clinical performance by adopting computer software applications to generate information items, such as drug alerts, other rule-based alerts, reminders, clinical guidelines and pathways. With the clinical decision support (CDS) tools, EHR system can effectively support the patient care and reduce the cost, time and errors in clinical workflow (Morris, 2005). Since the idea of CDS has first emerged, a lot of research endeavors have dedicated to the research and development of tools, methods and approaches for pattern discovery and predictive analysis. Some of those research studies explored the associations between diagnoses based on datasets of collected health records. IBM's HealthMiner and Molecular Concept Map (MCM) are two examples of data mining tools, which have been successfully applied in two different research studies to identify the related diagnoses among the health records (Mullins et al., 2005; Hanauer et al., 2009). Another research study investigates a method for selecting

* Corresponding author. Tel.: +852 34008596; fax: +852 23624365.
E-mail address: wing.chi.chan@inet.polyu.edu.hk (L.W.C. Chan).

0952-1976/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2011.05.013

the threshold of Chi-square test in analyzing co-occurring features, such as disease, finding and medication, of a large dataset of discharge summaries (Cao et al., 2007). These tools are very important and useful for the clinical experts to gain insights of disease patterns and linkages from the large EHR database. Besides CDS, clinical trials require the database retrieval of relevant cases for subject recruitment. A concept-based document search engine was implemented with an EHR system and evaluated through a retrospective clinical-epidemiological study targeting syphilis cases (Schulz et al., 2008). A research study proposed a sequence alignment strategy for finding patients with similar treatment histories in the temporal order of drugs given to the patients (Lee and Das, 2010). The approach could be useful for treatment cohort identification in clinical trials.

1.1. Motivation

As the methods for exploring disease patterns and linkages have been well developed, identifying similar health records based on similarity in disease pattern becomes a novel topic and growing research area in CDS. In some stages of clinical decision making processes, uncertainty exists and it is difficult to model the relevant medical domain knowledge in logical representations like rules. The similar health records retrieved using similarity matching algorithm are more concrete and convincing for medical experts to express their domain knowledge on particular disease patterns than the rule-based representation (Kong et al., 2008). For this reason, case-based reasoning approach presents similar health records to the clinicians as suggested solutions in many CDS applications, e.g. stress diagnosis based on finger temperature signals. To achieve the desired matching, the biomedical features were appropriately selected for those similarity matching algorithms based on the a-priori medical knowledge. For example, age, gender, room temperature, hours since meal, food and drink taken and finger temperature were considered as features for matching patients on similar stress level. Better goodness-of-fit can be attained when the parametric similarity measure is learnt through the adjustment of weights by domain experts or using the empirical data (Begum et al., 2009).

For matching patients with the same subclinical disorder like atherosclerosis, the etiology and underlying mechanism have not been completely known and clearly explained. The feature selection is not feasible as there are only a few variables that have significant association with the subclinical disorder and the association is usually very weak. An example of such variable is C-reactive protein (CRP). It was demonstrated in recent research studies that the elevation of CRP indicates increased risk of atherothrombotic events. However, it was shown in a large cross-sectional study that CRP is weakly associated with the manifestations of prevalent atherosclerosis, such as the intima-media thickness and ankle/brachial blood pressure index (Folsom et al., 2001). As all the available features are considered, a large dataset is required to train the parametric similarity measure. Otherwise, the trained measure may not be so accurate to match similar patient records. Therefore, non-parametric similarity measures are considered in this study.

Many classic similarity measures are non-parametric, including Euclidean distance (ED) and direction cosine (DC) (Qian et al., 2004). ED takes into the account the lengths of two vectors and the angle between them but DC depends on the deviation of vector direction only. Two similar health records retrieved based on DC could be very different in content. On the other hand, ED involves higher computational load than DC because ED considers all the vector components but DC needs not compute the zero-valued vector components, which are very often observed in the information retrieval applications. Combining the properties of ED and DC may yield a similarity measure addressing both precision and computational load issues. Further, the application of supervised machine learning approach could enhance the performance of these similarity measures through the relative weights of feature terms estimated by the training dataset.

1.2. Research questions

In this study, all the available disease terms in the health records are considered as the features and non-parametric similarity measure is used because most of the current EHR search engines are generally used for searching health records subject to any possible desired clinical terms and these tools were developed based on simple string matching of clinical terms. Thus, it is intuitive to hypothesize that a non-parametric similarity measure is a random scoring system in the agreement of a subclinical disorder between two patients. To cope with the randomness, the semantic relationships defined by medical ontology may help to estimate the closeness between patients of interest. Such kind of closeness, referred to as ‘‘clinical distance’’, was estimated by the semantic distance of the ‘‘minimal-cost’’ path in the ontology (Melton et al., 2006). As medical ontology can be used to align the synonymous and related clinical terms to unique concepts and form the feature space for the similarity measure, the likelihood of a common subclinical disorder found in two patients may be reflected by the similarity score based on medical ontology. To test the hypotheses for simple string matching and medical ontology, this study is aimed to examine whether the degree of similarity between two health records is associated with the subclinical disorder agreement between two patients, for the similarity measures based on simple string matching and medical ontology respectively. Further, a modified version of DC, referred to as mDC, is proposed in this paper and its performance in ranking the subclinical disorder agreement between two patients will be compared with that of ED and DC. This study is also aimed to examine whether the mDC outperforms the DC and ED in the estimation of patient similarity in terms of atherosclerosis. The implementation of the ontological model using machine learning approach will be also evaluated against the identification of subclinical disorder in the test dataset.

1.3. Core contribution and novelty of the work

Atherosclerosis is a major cause of cardiovascular disease (CVD). The American Heart Association (AHA) reported that no previous symptoms were observed in 50% of men and 64% of women who died suddenly of coronary heart disease (CHD) (Lloyd-Jones et al., 2009). Therefore, there is a need for search engine, which can retrieve similar cases from EHR system to provide concrete evidence for assessing the CVD risk of the asymptomatic individuals of interest. Once the above-mentioned research question is answered, it will bring out the potential of a medical-ontology-based search algorithm for identifying patients with similar CVD risk which is unlikely achieved by the current approach of simple string matching. From this study, the ontology-specific feature space and the derived similarity measure thereof will lead to a search engine, which will act as a simple, non-invasive and inexpensive tool to present similar cases to clinicians during consultation of asymptomatic patients. If the similar cases are already of CVD, the cases could be the evidence for predicting the future of the patients of interest and justifying the early intervention and prevention.

As a diabetic complication, atherosclerosis and its related information are not commonly documented in clinical practice. A study in southwest London showed that all of 17 participating