Chapter 16
A Preparation Framework for EHR Data to Construct CBR Case-Base

Shaker El-Sappagh
Mansoura University, Egypt

Alaa M. Riad
Mansoura University, Egypt

Mohammed Elmogy
Mansoura University, Egypt

Hosam Zaghloul
Mansoura University, Egypt

Farid A. Badria
Mansoura University, Egypt

ABSTRACT

Diabetes mellitus diagnosis is an experience-based problem. Case-Based Reasoning (CBR) is the first choice for these problems. CBR depends on the quality of its case-base structure and contents; however, building a case-base is a challenge. Electronic Health Record (EHR) data can be used as a starting point for building case-bases, but it needs a set of preparation steps. This chapter proposes an EHR-based case-base preparation framework. It has three phases: data-preparation, coding, and fuzzification. The first two phases will be discussed in this chapter using a diabetes diagnosis dataset collected from EHRs of 60 patients. The result is the case-base knowledge. The first phase uses some machine-learning algorithms for case-base data preparation. For encoding phase, we propose and apply an encoding methodology based on SNOMED-CT. We will build an OWL2 ontology from collected SNOMED-CT concepts. A CBR prototype has been designed, and results show enhancements to the diagnosis accuracy.

INTRODUCTION

Diabetes Mellitus (DM) is a serious disease. If it has not treated on time and properly, it can lead to serious complications including death. This makes diabetes one of the main priorities in medical science research, which in turn generates huge amounts of data. These data are transactional and distributed in the patient's EHR. An early diabetes diagnosis is the most critical step in diabetes management. The

DOI: 10.4018/978-1-5225-2229-4.ch016

diagnosis of diabetes is an ill-formed problem and depends on the physician experience. Case Based Reasoning (CBR) is considered as the most suitable Clinical Decision Support System (CDSS) for dealing with these problems where physicians share their experience (Richter and Weber, 2013; Blanco, 2013). Therefore, case-base creation is a challenging step. On the other hand, CBR is appealing in medical domains because a case-base already exists as the stored symptoms, medical history, physical examinations, lab tests, diagnoses, treatments, and outcomes for each patient (Andritsos et al., 2014). However, because clinical data are usually incomplete, inconsistent, and noisy, these data need a set of preparation steps before converted into CDSS knowledge (Abidi & Manickam, 2002). *The first step* is the data preprocessing stage that is applied to enhance data quality. The application of a set of machine learning algorithms improves the accuracy of CBR case retrieval algorithms. *The second step* is the coding stage that is used to represent the pre-processed data with standard coding terminology such as SNOMED CT (SCT) (Lee et al., 2013). We have proposed a diabetes diagnosis reference set from SCT version 2013 and modeled it in an OWL 2 ontology (El-Sappagh et al., 2014). This ontology is used to encode the unstructured (i.e. textual) contents of the case base knowledge base. Lack of standard data affects the accuracy of CDSS implementation (Ahmadian et al., 2011). Data standardization is critical for CBR systems for many reasons. The encoded knowledge supports: (1) the creation of distributed CBR systems; (2) the integration and interoperability between CDSS and EHR environment (Ahmadian et al., 2011); and (3) the creation of knowledge-intensive CBR systems. As a result, CBR supports semantic retrieval algorithms, and its intelligence is increased (Melton et al., 2006). Finally, *the third step* is the data fuzzification stage that is used to handle vague knowledge. Physicians always describe patients using vague terms, such as the sugar level is high, the patient has obese, and so on. Moreover, the patients often describe their conditions using imprecise terms. As Zadeh (2003) argued much of the knowledge that humans acquire through experience be perception-based and thus subject to imprecision and inaccuracy. Such knowledge, when not treated in some suitable way that can consider and convey its inherent imprecision, usually leads to reduced effectiveness of the knowledge-based systems that use it. Vagueness can be handled using fuzzy logic (Zadeh 2003), which has been used in diabetes diagnosis rule-based systems (Lee and Wang, 2011). Moreover, fuzzy logic has been integrated with CBR in hybrid systems (Abdul et al., 2014) and used for calculating the fuzzy similarity between cases (Khanum et al., 2009). However, in diabetes diagnosis domain, there are no studies in fuzzy CBR systems.

Authors in (Burnum, 1989; Weiner & Embi, 2009) stated that the introduction of health information technology like EHRs has not led to improvements in the quality of the data being recorded, but rather to the recording of a greater quantity of bad data. As a result, Lei (1991) has proposed what he called the first law of informatics: "data shall be used only for the purpose for which they were collected." In the same time, EHR contains all the current and history of medical data of the patient. These data can be used as a complete source for building the CBR's case-base (Abidi & Manickam, 2002). The quality of CBR is based on the quality of case-base content (Andritsos, 2014). EHR data quality measurement and improvement must be an essential step in using its data in CDSS's knowledge base (Abidi & Manickam, 2002). As a result, data preprocessing steps are the first and the foremost to improve the accuracy of CBR systems (Borges et al., 2012). By focusing on DM diagnosis, its medical dataset is seldom complete (Jayalskshmi & Santhakumaran, 2010). Moreover, because diabetes is a lifelong disease, even data available for an individual patient may be massive and complicated to interpret. Data preprocessing steps include deleting of low-quality rows and columns, feature selection, feature mining, integration, transformation (i.e. normalization and discretization), data cleaning, feature weighting, etc. (Begum et al., 2010). An example of a system focusing on feature mining is the dietary counseling system by Wu et al. (2004).