EHR Data Preparation for Case Based Reasoning Construction

Shaker El-Sappagh[1,2], Mohammed Elmogy[1], A. M. Riad[1], Hosam Zaghlol[3], and Farid A. Badria[4]

[1] Faculty of Computers & Information, Mansoura University, Egypt
[2] Faculty of Computers & Information, Minia University, Egypt
[3] Faculty of Medicine, Mansoura University, Mansoura, Egypt
[4] Pharmacognosy Department, Faculty of Pharmacy, Mansoura University, Mansoura, Egypt

Abstract. Case Based Reasoning (CBR) is the first choice in experience-based problems as diagnosis. However, building a case base for CBR is a challenging. Electronic Health Record (EHR) data can provide a starting point for building case base, but it needs a set of preprocessing steps. In this paper, we propose a case-base preparation framework for CBR systems. This framework consists of three main phases including data preparation, fuzzification, and coding. This paper will focus only on the data-preprocessing phase to prepare the EHR database as a knowledge source for CBR cases. It will use many machine-learning algorithms for feature selection and weighing, normalization, and others. As a case study, we will apply these algorithms on diabetes diagnosis data set. To check the effect of data preparation steps, a CBR prototype will being designed for diabetes diagnosis and prediction of its complications as kidney failure. The results show an enhancement to the case retrieval process of the implemented CBR system.

Keywords : case based reasoning, data preprocessing, diabetes diagnosis, clinical decision support system, electronic health record, and case-base knowledge.

1 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. Early diabetes diagnosis is the most critical step in diabetes management. The 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 [1, 2]. 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 storing symptoms, medical history, physical examinations, lab tests, diagnoses, treatments, and outcomes for each patient [3]. However, because clinical data are usually incomplete,

A.E. Hassanien et al. (Eds.): AMLTA 2014, CCIS 488, pp. 483–497, 2014.

inconsistent, and noisy, these data need a set of preparation steps before converted into CDSS knowledge [4]. The first step is the data preprocessing stage that is applied to enhance data quality. The second step is the data fuzzification stage that is used to handle vague knowledge. Finally, the third step is the coding stage that is used to represent the processed data with standard coding systems such as SNOMED CT [5].

Authors in [6, 7] 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 [8] 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 [4]. The quality of CBR is based on the quality of case-base content [3]. EHR data quality measurement and improvement must be an essential step before using its data in CDSS's knowledge base [4]. As a result, data preprocessing steps are the first and the foremost to improve the accuracy of CBR systems [9]. By focusing on DM diagnosis, its medical dataset is seldom complete [10]. Moreover, because diabetes is a lifelong disease, even data available for an individual patient may be massive and difficult 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. [11]. An example of a system focusing on feature mining is the dietary counseling system by Wu et al. [12]. Jagannathan and Petrovic [13] have concluded that missing values in the case-base pose a common and serious problem that impairs the performance of the system, and they have provided an imputation method to deal with missing value. However, they have only handled missing values of some attributes, and they have left others. Floyd et al. [14] have concluded that applying preprocessing techniques to a case-base as feature selection can increase the performance of a CBR system.

Case retrieval is the most important phase in CBR system. However, it is mainly depend on the types of case-base knowledge and its quality. All existing case retrieval algorithms depend on one type of knowledge for retrieval [43]. We will combine the most important three techniques for data preparation to enhance the case retrieval process especially for medical domain. In this paper, our proposed framework will result in cleaned, normalized, fuzzified, and encoded knowledge. These types of knowledge will support different queries and different types of similarity algorithms, which improve the retrieval phase of CBR system. Diabetes diagnosis is used as a case study for applying this framework. There are not CBR systems that utilize machine-learning algorithms to prepare case-base knowledge. Moreover, most of CBR systems for diabetes diagnosis have used a diabetes specific data set. However, diabetes as a chronic disease results in many other diseases as nephropathy, retinopathy, neuropathy, heart diseases, stroke, and others [42]. In our data set, patient is described by 70 different features, as shown in Table 1. These features link diabetes with other diseases, such as cancer, kidney disease, and liver diseases. A CBR-based CDSS for diabetes diagnosis will be designed using myCBR 3 protégé plug-in [36]. The diagnosis will be the patient's diabetes status plus his future conditions of having other

diseases, such as cancers. The paper is organized as follows. Section 2 provides related works, section 3 provides a description of our diabetes data set, section 4 provides the proposed preparation framework, section 5 provides a case study for CBR system, and finally section 6 provides the conclusion and future works.

2 Related Work

Data quality of EHR should be measured before using it as a knowledge source. Weiskopf and Weng [15] have determined five dimensions to measure the quality of EHR data. Klompas et al. [16] has asserted that EHR data can improve diabetes management even if its raw data quality is low. For achieving data quality, a preprocessing or data preparation step is critical for any knowledge based CDSS system [17]. For example, the case retrieval algorithms, such as nearest neighbor, require data cleaning and normalization steps [13, 18]. Low quality data need handling in CBR. For example, Xie et al. [19] have handled missing values and unmatched features in case retrieval algorithm. Guessouma et al. [20] have proposed five approaches for managing missing data problem in a medical CBR system. The data preprocessing steps include data cleaning [13], data transformation [10], feature mining and selection [21], etc. Especially for CBR, the conversion of database structure to case-base structure is a critical step [4]. Database record is similar to case-base case, but transformation from generic EHR to specialized case-base is not a straightforward mapping of attributes. In order to change one of simple database records into a case, it is required to associate an experience to the record [22]. Abidi et al. [4] have assumed that the structure of case-base is defined in advance. They have mapped the structure then contents of EHR to case-base. However, this assumption is not realistic. The structure of the case-base depends on the structure and contents of EHR, and it must be inferred from it. As EHR contains patient raw chronicle data, a temporal abstraction preparation step is critical to aggregate and provide trends from patient data [23]. Moreover, features weights must be specified for case retrieval algorithms [1]. It can be determined manually by domain expert and CPGs [20] or automatically using machine learning algorithms [24], e.g. neural network [4]. The choice of case features that best distinguish classes of instances has a large impact on the similarity measure [22], and it improves the performance and decrease the complexity [3]. This process can be done automatically [25] (using techniques as information gain [26] and Relief [27]) or manually [28, 29] according to domain expert or CPGs. Kotsiantis [30] has surveyed data preprocessing algorithms for each step. Han et al. [31] have proposed the preprocessing steps for a DM data set using RapidMiner [37]. There is no single sequence of data pre-processing algorithms with the best performance [30]. As a result, data preprocessing is a set of not ordered steps [17]. It can be an inherent component of a CBR system, or it can be performed as a preprocessing step. For example, eXiT*CBR [32] system incorporate basic preprocessing steps as discretization, normalization and feature selection techniques. However, the complete list of needed preprocessing steps is different according to the nature of data and the CBR system purpose [3]. As a result, this paper will provide the preprocessing step as a separate phase before CBR system processes.

3 Data Set Description

The paper uses a data set from diagnostic biochemical lab, AutoLab of Mansoura institution, Mansoura University, Mansoura, Egypt. This data was collected in the period from January 2010 through August 2013. The control subjects were healthy and recruited from the diagnostic biochemical lab and were matched by age, sex and ethnicity to the case subjects. The eligibility criteria for controls were the same as those for patients, except for having a cancer diagnosis. A short structured questionnaire was used to screen for potential controls based on the eligibility criteria. Analysis of the answers received on the short questionnaire indicated that 80% of those questioned agreed to participate in clinical research. A total of 67 eligible subjects were ascertained in the current study. However, 7 control subjects were excluded due to limited blood samples for testing AFP. Blood samples (5 mL) were taken, centrifuged, and the serum separated and stored at 220uC until analyzed. Serum samples were assayed for AFP by enzyme-linked immunosorbent assay with commercial kits (Abbott, North Chicago, IL), transferase (ALT) and aspartate aminotransferase (AST), with an auto-analyzer (Hitachi Model 736, Japan) and commercial kits.

The problem features include: Demographics (Residence, Occupation, Gender, Age, BMI); Lab tests (HbA1C, 2h PG, FPG); Hematological Profile (e.g. Prothrombin INR, Red cell count, Hemoglobin, Haematocrit (PCV), MCV, MCH, MCHC, Platelet count, White cell count, Basophils, Lymphocytes, Monocytes, Eosinophils); Symptoms (Urination frequency, Vision, Thirst, Hunger, Fatigue); Kidney Function Lab tests (Serum Potassium, Serum Urea, Serum Uric acid, Serum Creatinine, Serum Sodium); Lipid Profile (LDL cholesterol, Total cholesterol, Triglycerides, HDL cholesterol); Tumor Markers (FERRITIN, AFP Serum, CA-125); Urine Analysis (Chemical Examination (Protein, Blood, Bilirubin, Glucose, Ketones, Urolibingen), Microscopic Examination (Pus, RBcs, Crystals)); Liver Function Tests (S_Albumin, Total Protein, Total Bilirubin, Direct Bilirubin, SGOT (AST), SGPT (ALT), Alk_Phosphatase, γ GT); Females History (Amenorrhea, Birth, Dysmenorrhea). Solution features include: Diabetes Diagnosis; Nephropathy check; Hypercholestremia check; Tumor Markers; Liver problem; Radiological Diagnosis (Glomerulonephrinitis, Shrunken kidney, HCC, HCV, Ovarian cancer, Fatty Liver, Splenomegaly). Table 1 is a sample of the tested features, and we have selected a unified Unit of Measurement (UoM) for each lab test. All features' values are converted to the selected UoM.

Table 1. Patient features (Data type:N=Numerical, C=Categorical, and O=Ordinal}.

Feature type	Feature name	Data type	Normal Range	UoM	Min-Mean-Max
Demographics	BMI	N	18.5 - 25	kg/m^2	20-33.117-45
Diabetes Lab Tests	HbA1C	N	<=5	mmol/L	5-6.373-7.4
Hematological Profile	Hemoglobin	N	12 - 16	g/dL	9.8-12.332-13.4
	White cell count	N	4 - 11	10^3/cmm	6-8.055-9.2
Symptoms	Urination frequency	O	-	-	-
Kidney Function Lab tests	Serum Uric acid	N	3.0 - 7.0	mg/dL	3-4.237-7.9
	Serum Creatinine	N	0.7 - 1.4	mg/dL	0.9-1.35-3.6
Lipid Profile	LDL cholesterol	N	0 - 130	mg/dL	50-94.917-170
Tumor Markers	FERRITIN	N	28 - 397	ng/mL	-
Liver Fun. Tests	S. Albumin	N	3.5 – 5.0	g/dL	1.9-4.082-5.4
Females History	Amenorrhea	O	-	-	-
Diagnosis	Diabetes Diagnosis	C	-	-	-
	Nephropathy check	C	-	-	-
	Hypercholestremia check	C	-	-	-
	Tumor Markers	C	-	-	-
	Liver problem	C	-	-	-

Table 2 show a sample of 11 case with all features describing diabetic patients. These raw data will be prepared as a case base knowledge for diabetes diagnosis CDSS based on CBR technique. Our case base will contain 60 cases.

Table 2. A sample of 11 cases from our raw data set before preprocessing

Case	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10	Case 11
BMI	20	23	31	32	40	45	24	31	29	28	29
Gender	M	F	F	M	M	F	F	F	M	F	F
Age	39	47	32	35	53	43	41	32	35	53	60
Vision	blurred	Allergy /redness	Non	Non	Allergy /retinopathy	retinopathy	non	Aller-gy/redness	Non	Non	Allergy/ retinopathy
Fatigue	Non	Non	+	+	+	+	+	+	++	+	+
Hunger	Non	Non	+	+	++	+	Non	Non	+	+	Non
Thirst	Non	+	+	+	+	++	+	Non	+	+	Non
Urination Freq.	Non	+	Non	Non	++	+++	Non	Non	++	++	Non
Residence	urban	Rural	Rural	urban	urban	urban	urban	urban	urban	urban	Rural
Occupation	Teacher	Merchant	Farmer	Engineer	worker	Pharmacist	Doctor	Non	Worker	Merchant	Teacher
2hPG	180	190	200	210	220	230	165	225	175	185	230
FPG	120	125	130	135	140	150	118	148	117	121	150
HbA1C	6.4	6.5	6.6	6.7	6.9	7.1	6.1	6.8	6.2	6.3	7.1
Prothrombin (INR)	1.3	1.1	1	1.1	1.4	1	1.4	1.1	1	1.3	1.4
Basophils	0	1.1	0	1.2	0.5	1	1.5	1.2	0.5	1	1
Eosinophils	1	2	3.4	1.8	1.4	2.4	1.2	1.8	1.4	2.4	2.4
Monocytes	2	3.8	2.5	3.3	2	4	2.2	3.3	2	4	4
Lymphocytes	25	29	24.7	26.8	23.2	27	21.2	26.8	23.2	27	27
White cell count	7.9	9.2	8.8	8.4	7	8.9	6	8.4	7	8.9	8.9
Platelet count	232	2000	195.4	210	198	210	192	210	198	210	210
MCHC	32.8	37.8	41.7	38.7	30.8	32.8	30	41.7	38.7	37.8	30.8
MCH	23.4	28.4	29.4	28.4	21.4	23.4	21.3	29.4	28.4	28.4	21.4
MCV	71.3	74.5	76.4	74.5	70	71.3	70	76.4	74.5	74.5	70
Haematocrit	34.8	36.8	34.5	36.8	34.3	36.8	31.1	34.5	36.8	36.8	34.3
Hbg	11.4	12.8	13.4	12.8	10.4	12.5	9.8	13.4	12.8	12.8	10.4
Red cell count	4.88	5.4	5.88	5.4	4.1	5.2	3.8	5.88	5.4	5.4	4.1
Serum Potassium	3.8	2.4	4.1	4.3	2.9	4	3.9	3.5	4.3	4.1	3.9
Serum Sodium	135	149	135	134	152	134	135	135	134	135	135
Serum Creatinin	2.6	1.9	1	0.9	2.4	1.1	1	1	0.9	1	1
Serum Uric acid	7.8	5.2	3.4	4.9	7.9	3	3.4	3.4	4.9	3.1	3.4
Serum Urea	56	47	18	18	52	28	17	17	28	19	17
LDL cholesterol	90	85	55	67	165	81	79	89	165	167	55
HDL cholesterol	51	55	60	61	35	65	65	65	38	38	65
Triglycerides	158	141	152	150	180	142	134	101	181	167	111
Total cholesterol	200	200	200	200	240	200	200	200	243	255	200
FERRITIN	Normal	Normal	Normal	Normal	Normal	Normal	Normal	Normal	Normal	Normal	Normal
AFP Serum	Normal	Normal	Normal	Normal	Normal	Normal	Normal	Abnormal	Normal	Normal	Abnormal
CA-125	Normal	Normal	Normal	Normal	Normal	Abnormal	Normal	Normal	Normal	Normal	Normal
Crystals	+	Nil	Nil	Nil	++	Nil	Nil	Nil	Nil	Nil	Nil
RBcs	+	Nil	Nil	Nil	++	+++	Nil	Nil	Nil	Nil	+
M.Pus	++	Nil	Nil	Nil	+++	+++	Nil	Nil	Nil	Nil	+
Urolibingen		Nil	Nil	Nil		Nil	Nil	Nil	Nil	Nil	Nil
Ketones	++	+	Nil	Nil	++	+++	Nil	Nil	Nil	Nil	++
Glucose	+++	++	Nil	Nil	+++	+++	Nil	Nil	Nil	Nil	+++
Bilirubin			Nil	Nil		Nil	Nil	Nil	Nil	Nil	Nil
Blood		+	Nil	Nil	++	Nil	Nil	Nil	Nil	Nil	+
Protein	++	++	Nil	Nil	+++	Nil	Nil	Nil	Nil	Nil	+
Albumin	4.5	4.1	5	5.4	4.5	4.4	2.4	1.9	4.1	5	2.8
Total protein	4.7	3.2	6.4	3.8	3.1	4.1	8.1	8.2	3.2	6.4	7.5
γ GT	22	20	18	27	25	28	64	98	20	18	68
Alk. phosphatase	250	189	300	210	170	184	350	289	189	300	210
SGPT (ALT)	35	45	42	40	35	42	110	82	45	42	123
SGOT(AST)	40	41	39	35	40	40	98	78	39	35	145
Direct bilirubin	0.3	0.4	0.3	0.5	0.3	0.5	1.4	1.3	0.3	0.4	1.6
Total bilirubin	1	1.1	1	1.2	0.9	1	2.8	3	1.1	1	2.2
Birth	Normal	Normal	Normal	Normal	Normal	twins	Normal	Overweight baby	Normal	Normal	Normal
dysmenorrhea	Normal	Normal	++	Normal	Normal	Normal	Normal	Normal	Normal	Normal	Normal
Amenorrhea	Normal	Normal	+	Normal	Normal	Normal	Normal	Normal	Normal	Normal	Normal
Diabetes Diagnosis	Prediabetic	Diabetic '	Diabetic gestational'	Diabetic	Diabetic	Diabetic	Prediabetic	Diabetic gestational'	Prediabetic	Prediabetic	Diabetic
Nephropathy Diagnosis	Nephropathy	Nephropathy	Normal	Normal	Nephropathy	Normal	Normal	Normal	Normal	Normal	Normal
Hypercholestremia Diagnosis	Normal	Normal	Normal	Normal	Hypercholestre mia	Normal	Normal	Normal	Hypercholestre mia	Hypercholest remia	Normal
Cancer Type	Normal	Normal	Normal	Normal	Normal	Ovarian cancer	Normal	Hepatocellular carcinoma	Normal	Normal	Hepatocellular carcinoma
Liver Diagnosis	Normal	Normal	Normal	Normal	Normal	Normal	HCV	HCC	Normal	Normal	HCV
Glomerulonephrinitis	No	Yes	No	No	Yes	No	No	No	No	No	No
shrunken kidney	No	Yes	No	No	Yes	No	No	No	No	No	No
HCC	No	No	No	No	No	No	No	Yes	No	No	No
HCV	No	No	No	No	No	No	Yes	No	No	No	Yes
Ovarian cancer	No	No	No	No	No	Yes	No	No	No	No	No
Fatty Liver	No	No	No	No	No	No	No	No	No	No	No
Splenomegaly	No	No	No	No	No	No	Yes	No	No	No	No

4 Case-Base Preparation Framework

In this section, we propose a case-base preparation framework for extracting a diabetes diagnosis case-base from EHR databases, as shown in Fig. 1. This framework discusses the conversion of both structure and content of EHR database to a derived

case-base structure and content. There are three sequential phases for creating a case-base from EHR including data preprocessing, data fuzzification, and data encoding. We have proposed, in other work, a standard data model based on HL7 RIM [38] to standardize the structure of case base. This data model will be utilized to prepare the case-base structure. Moreover, we have created an OWL 2 ontology for diabetes concepts from SNOMED CT to be used for coding and standardizing the case-base contents. Encoding phase is performed to standardize the textual contents of case-base according to standard medical ontology [5]. Because of space restrictions, the data preparation phase will be discussed in this work, and the other two phases will be considered in future works. Because it is the general case, we will assume that EHRs (the case sources) are in relational database format [33], and do not use any coding mechanisms (e.g. the data are in primitive types only like numerical, textual, etc.)

4.1 Data Extraction, Integration, and Anonymization

EHR contains medical, administration, financial, security, and privacy information. By concentrating on medical data, it contains heterogeneous data about patient, related to his lifelong health states. Moreover, that information structure and representation format of EHR varies across different HISs. Interoperability of EHRs is out of scope. We need to collect only data fields related to DM diagnosis from all of patients EHRs into a coherent data store. According to our domain expert and Clinical Practice Guidelines (CPGs), the list of diagnostic data elements are identified including lab tests, symptoms, and others. These data elements are extracted from multiple sources (i.e. EHRs). These sources require object matching and schema integration strategies. Data integration can help reducing and avoiding redundancies and inconsistencies in the resulting data set. We will follow a manual process for schema integration, i.e. attributes names conflict. Object matching requires attribute tuple redundancy detection and prevention. Anonymization is a privacy preservation step. We have removed all patient personal details, and the paper will depend on an artificial identifier for each case.

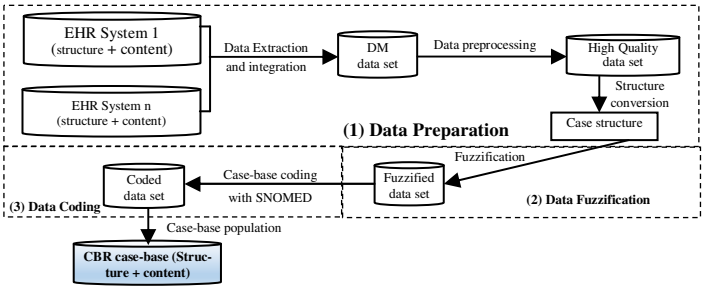

Fig. 1. Case-base preparation phases

4.2 Data Preprocessing Steps

This paper focus on the representation of case-base in attribute-value tables, because it is one of the most common kind of case representation in CBR. Moreover, the attributes types will be nominal and numerical only with no object and concept types. Currently available options in our CBR are: feature weighting, feature selection, outlier detection and removal, handling missing values, categorical feature coding, and others. These steps are performed sequentially on the raw case-base data to produce a new high quality case base. Fig. 2 shows that case-base pre-processing steps are sequential, where P_i is a pre-processing step.

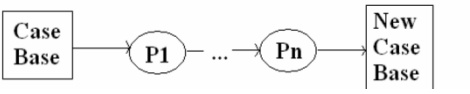

Fig. 2. The case-base pre-processing steps

4.2.1 Handling Missing Values and Categorical Features Coding

The effect of incomplete data on retrieval performance is an important issue because it affects both precision and recall of retrieval algorithm [34]. There are so many methods for handling missing values. In our case, we will remove: (1) all cases with more than or equal to 50% of missing features, (2) all features with more than or equal to 50% of missing values. Missing values will be filled using the k-NN algorithm [41], where the corresponding feature of the most similar case will fill the missed feature. The features data types are numerical, ordinal, and nominal.

Table 3. Categorical and diagnosis features codes

Categorical Feature		Coded Values	Original value			# of cases
Urination Frequency		0	Non (Normal) *i.e. 3-5 times urination per day*			42
		1	*+ i.e. 6-8 times urination per day*			4
		2	*++ i.e. 9-10 times urination per day*			11
		3	*+++ i.e. more than 10 times urination per day*			3
CA-125		0 (i.e. Normal)	Numerical values			57
		1 (i.e. Abnormal)	Numerical values			3
FERRITIN		0 (i.e. Normal)	Numerical values			58
		1 (i.e. Abnormal)	Numerical values			2
AFP Serum		0 (i.e. Normal)	Numerical values			56
		1 (i.e. Abnormal)	Numerical values			4
Urine Analysis		0 (i.e. Nil)	1 (i.e. +)	2 (i.e. ++)	3 (i.e. +++)	
	Protein	50	1	7	2	
	RBcs	49	9	1	1	
	Crystals	55	4	1	0	
Diabetes Diagnosis		0	Normal			8
		1	Prediabetic			19
		2	Prediabetic/Gestational			1
		3	Diabetic			29
		4	Diabetic/Gestational			3
Nephropathy Check		0= Normal	Empty			49
		1= Abnormal	Nephropathy			11
Hypercholestremia Check		0= Normal	Empty			46
		1= Abnormal	Hypercholestremia			14
Cancer Type		0	Empty (i.e. Normal)			52
		1	Ovarian cancer (Preneoplasm)			3
		2	Liver cirrhosis			1
		3	HCC (liver cancer or Hepatocellular carcinoma)			4
Liver Problem		0	Normal			50
		1	Fatty (bright) liver			3
		2	HCV (virus C)			3
		3	HCC (liver cancer)			4

Numerical features (i.e. lab tests) will not be coded. As a special case of numerical features, FERRITIN, AFP Serum, and CA-125 will be coded as Normal and Abnormal. On the other hand, all categorical and ordinal features (e.g. patient symptoms) will be coded. Moreover, the diagnosis features (the case solution) which include diabetes diagnosis, Nephropathy check, Hypercholestremia check, kidney problems, liver problems, and Radiological diagnoses will also be coded. RapidMiner Studio 6.0 has done the coding process, using Discretize Operator. This operator discretizes the selected attributes into user-specified classes using a simple mapping process. Table 3 shows a sample of the coded features.

4.2.2 Feature Selection

A sufficient and appropriate description for the problem is necessary to describe it in a case-base form [35]. The collected features for patients are huge, and they have different levels of importance for determining the diagnosis of diabetic patient. Initial feature vector is collected from domain expert, EHR database schema and diabetes diagnosis CPGs, as discusses in section 3. The large number of features affects the performance of the case retrieval algorithm. If extraneous and misleading features are removed, the similarity measures can retrieve cases that are more useful. As a result, feature selection algorithms must be applied on the collected features to determine the most important ones. There are two types of feature selection methods: Wrapper methods and Filter methods [39]. As the two methods are complementary, we will try algorithms from both approaches. We have applied a set of machine learning algorithms [39], and we have compared their results. For example, in WEKA, we have examined K-NN classifier, Naïve Bayes, C4.5 decision tree, and fuzzy rough feature selection. Moreover, in RapidMiner, we have examined rule induction technique in backward elimination optimization technique, rule induction technique in forward selection optimization technique, k-NN technique in backward elimination optimization technique, and decision tree technique in genetic algorithm optimization. For space restrictions, we will not discuss any details about these algorithms. We can range the most important features repeated in feature selection algorithms. The most important feature is HbA1c, and it will have the highest weight in CBR system. Moreover, it can be noticed that feature selection algorithms have selected overlapped sets of features. As a result, we will combine the collected features from these algorithms to participate in case representation in our CBR system.

4.2.3 Feature Weight Assignment

Feature weighting process is used to improve the performance of the case retrieval algorithm. Weights can be attached to cases, so that cases considered more important for a given application when have higher weights. Weight vectors can also be assigned to description variables: one can either use the same weight vector for all cases, or assign individual weight vectors to each case. As a result, high significant attributes will receive higher weights. This paper calculates, using machine learning algorithms and tools, a single weights vector for features of the entire case-base. Feature weighting will be calculated by using variety of algorithms. In Table 4, we have calculated features weights using the following algorithms [37, 40]: A= Genetic

Algorithm + decision tree, B= Genetic Algorithm + rule induction, C= Particle swarm optimization, D= Information Gain technique, E= Correlation Technique. Because some features have been selected in feature selection, but they have weight = zero in feature weighting algorithms, we will take the maximum of these weights as the feature weight vector in our CBR system. Table 3 shows a sample of features weights that are vales ∈ [0, 1]. All selected features in previous section will have weights.

Table 4. Features weights

	A	B	C	D	E	Maximum
2hPG	0.060	0.001	1.000	0.894	0.038	**1**
Birth	0.128	0.551	0.396	0.206	0.248	**0.551**
CA125	0.515	0.011	0.000	0.043	0.298	**0.515**
Blood	0.124	0.241	0.000	0.110	0.410	**0.410**
Ketones	0.254	0.129	0.000	0.290	0.736	**0.736**
D. bilirubin	0.000	0.004	0.000	0.082	0.172	**0.172**
FERRITIN	0.312	0.204	0.000	0.024	0.305	**0.312**
Hunger	0.377	0.202	0.929	0.074	0.308	**0.929**
HbA1C	0.030	0.025	0.370	1.000	0.049	**1**
Hemoglobin	0.092	0.176	0.619	0.157	0.699	**0.699**
Platelet count	0.020	0.118	1.000	0.111	0.422	**1**
Prothrombin INR	0.134	0.057	0.056	0.051	0.689	**0.689**
Red cell count	0.047	0.089	0.897	0.157	0.745	**0.897**
Residence	0.063	0.391	1.000	0.031	0.520	**1**
SGOT_AST	0.150	0.106	0.000	0.070	0.271	**0.271**
SGPT_ALT	0.061	0.000	0.000	0.059	0.311	**0.311**
S. Potassium	0.099	0.165	0.047	0.077	0.746	**0.746**
S. Sodium	0.104	0.144	0.095	0.098	0.363	**0.363**
S. Uric acid	0.000	0.136	1.000	0.087	0.522	**1**
Triglycerides	0.014	0.127	0.000	0.101	0.790	**0.790**
White cell count	0.096	0.020	1.000	0.134	0.221	**1**

4.2.4 Outlier Detection and Handling

Retrieval performance in CBR is likely to be adversely affected if noise is presented in the case library. While many learning algorithms can be modified to be noise tolerant, it is difficult to make instance-based learning (IBL) algorithms such as k-nearest neighbor (k-NN) algorithm or case-based reasoning (CBR) robust against noise. Moreover, outliers will affect the normalization process in the next section. In our data set, we have checked for outliers using RapidMiner's Detect Outlier (Distances) operator. This operator identifies n outliers in the given dataset based on the distance to their k nearest neighbors. We have found that all outlier cases are tightly affected by extreme values of some specific features, and these values are replaced by our domain experts.

4.2.5 Normalization of Numerical Attributes

Most CBR retrieval algorithms (e.g. k-NN) are depending on distance similarity functions as Euclidean distance. Therefore, all attributes should have the same scale for a fair comparison between them. Normalization of the data is very important when dealing with attributes of different units and scales. Regarding normalization of numerical attributes, a key issue is having all the local similarity measures in the same scale [0, 1] to combine them at the global similarity measure. There are many normalization techniques as Z-score and Min-Max. RapidMiner has normalize operator, which normalize all of the numerical features using Min-Max technique. All features are normalized in specific [C, D] range using Eq.1 where A is the old value, B is the normalized value. The used range in our case is [0.0, 1.0].

$$B = \left(\frac{A - \text{minimum value of } A}{\text{maximum value of } A - \text{minimum value of } A}\right) * (D - C) + C \quad (1)$$

After the previous preprocessing steps, our normalized data set is now ready for building the diabetes diagnosis case-base. In the following section, we will build a CBR system prototype using myCBR3 protégé plugin [36]. The myCBR is an open-source similarity-based retrieval tool, and it comes in three forms: protégé plugin, workbench, and software development kit (SDK). In this paper, we will use the first form.

5 Case Study

To check the effects of data preprocessing on the quality of CBR results, we will provide a prototype of a CBR system using myCBR framework [36]. The following subsections describe the developing phase of CBR.

5.1 Case-Base Formulation

We have a case-base containing 60 cases in the form: $CaseBase = \{case_1, case_2 \ldots case_{60}\}$, where $case_i = case\ (F_i, S_i)$, present the i^{th} case of database. Each case has ID. $F_i = (f_{i1}, f_{i2} \ldots f_{in})$ presents the problem description features of case i, and f_{in} present the n^{th} feature of case C_i. These are 58 features includes patient's *lab tests* as HbA1c and *symptoms* as Thirst. $S_i = (s_{i1}, s_{i2} \ldots s_{in})$ presents the solution sets of case C_i, whereas s_{in} presents the n^{th} solution of case C_i. Each solution has 12 features and determines the diabetes diagnosis of a patient plus his probability of having a kidney problem (e.g. Glomerulonephrinitis, shrunken kidney, etc.), liver problem (e.g. Fatty liver, HCV, HCC, etc.), Hypercholestremia, and Nephropathy. The *myCBR* supports case representation by CSV data file in the form of attribute-value using CSV data import module. Fig. 3 shows the myCBR screen after importing the data set into a new class Patient that will be used as query and case values for retrieval step.

5.2 Modeling Similarity Measures

The k-NN algorithm is used for case retrieval. It follows the local-global approach that divides the similarity definition into a local similarity measures for each attribute, weight for each attribute, and a global similarity measure for calculating the similarity of cases. Different types of features have different similarity comparison methods. For attributes of the data type float, integer, and ordinal, we used distance functions. We used similarity tables for symbolic value ranges. In our case, if each case has n attributes then the similarity between query q and case c is calculated using Eqs. 2, 3, 4, as shown in Fig. 4.

$$Dist\,(q,c) = \sqrt{\sum_{i=1}^{n} w_i \times D_i(q_i, c_i)} \tag{2}$$

$$D_i(q_i, c_i) = \begin{cases} (q_i - c_i)^2, & \text{if } q_i \text{ and } c_i \text{ are numerical or ordinal} \\ 1 \quad , & \text{if } q_i \text{ and } c_i \text{ are nominal and } q_i \neq c_i \\ 0 \quad , & \text{if } q_i \text{ and } c_i \text{ are nominal and } q_i = c_i \end{cases} \tag{3}$$

$$Sim(q,c) = f\big(Dist\,(q,c)\big), \qquad \text{e.g. } Sim(q,c) = \frac{1}{1+\propto Dist\,(q,c)} \tag{4}$$

Where D_i and w_i denote the local similarity measure and the weight of attribute i, respectively. Sim is the similarity between q and c. $Dist$ represents the weighted Euclidian distance between two cases which is special case of Minkowski Distance (MD). The choice of the parameter p depends on the importance we give to the differences.

$$MD\,(q,c) = \left(\sum_{i=1}^{n} |q_i - c_i|^p\right)^{\frac{1}{p}} \tag{5}$$

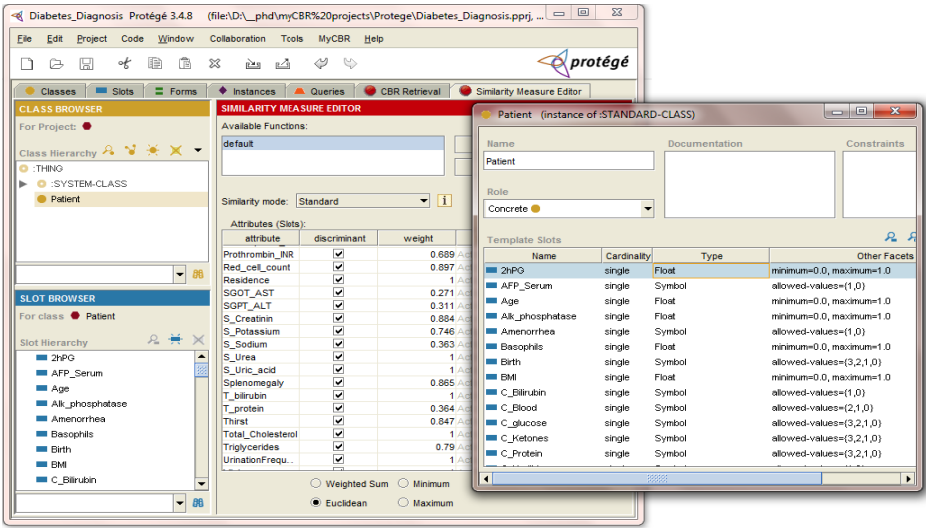

Fig. 3. The patient case data representation in myCBR

The weight values of the case Id and Class features are zero. The myCBR support the modification of the default similarity functions for nominal and numerical feature. Moreover, it supports the definition of similarity between UNKNOWN and UNDEFINED values, and it allows the creation of other special values.

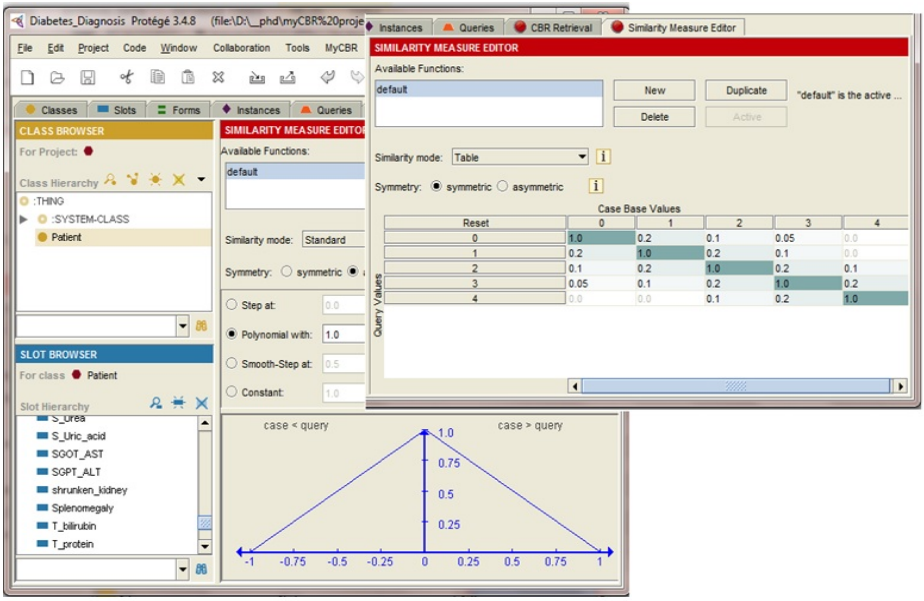

Fig. 4. The features local similarity functions

5.3 Testing of Retrieval Functionality

Case retrieval is the most critical step to test CBR functionality, and the whole purpose of our proposed framework is to improve it. The definition of an optimal similarity measure is often a difficult and tricky task. It requires repeatedly testing and fine-tuning. We have tested the 60 cases that are exist in the case base. The physician query will be normalized and coded before entered to the CBR system. We have tested our framework with a list of new cases. Cases descriptions will be entered to the system as queries, and the output will be the patient diabetes diagnosis including normal, prediabetic, diabetic, etc. Moreover, the systems will provide the patient probability to produce other complications as kidney failure. Retrieved cases are sorted by the level of similarity to the query. Fig. 5 shows one query after retrieving the most similar cases using CBR Retrieval tab in protégé. Moreover, the Queries tab in protégé can be used to build queries. We have measured the retrieval accuracy of the system. The accuracy means the ability of the system to retrieve the right or the similar case. The applied preprocessing steps have increased the accuracy of the CBR system to find the most suitable case. The value of k has not determine because myCBR framework will return all the case with their similarity level ranging from 100 (i.e. exact similarity) to 0 (i.e. not similar). All tested cases are retrieved with 100% accuracy. The retrieved case may need some adaptation to generate the final solution, but this task is out of scope.

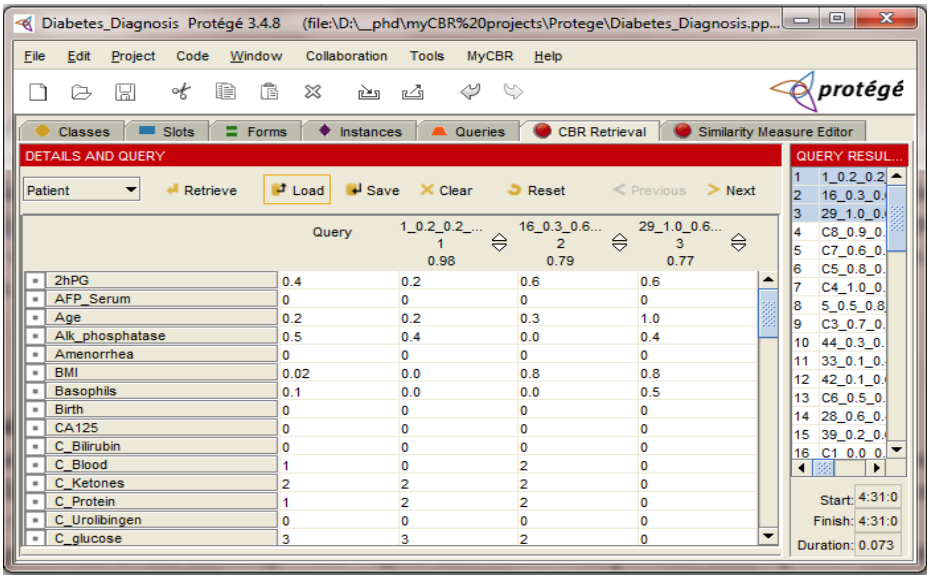

Fig. 5. The case-base query retrieval

6 Conclusion

In this paper, we have proposed a case-base preparation framework. The paper concentrated on the data preprocessing steps including missing values handling, feature selection, feature weighing, outlier detection, and normalization. Multiple machine learning algorithms have been applied to preprocess the case-base data. A CBR system has been implemented as a case study on diabetes diagnosis data set, and the results has been recorded. The preparation of EHR data enhances the accuracy of the retrieval phase of our CBR system. In our future works, we will complete the implementation of our case-base preparation framework including the fuzzification of the case base, implementation of fuzzy similarity technique and codifying of the case-base using a standardized ontology as SNOMED CT, ICD, or UMLS. Moreover, we will increase our case-base size with new cases to enhance the accuracy the CBR system decision.

References

1. Richter, M., Weber, R.: Case-Based Reasoning. Springer, Heidelberg (2013)
2. Blanco, X., Rodríguez, S., Corchado, J.M., Zato, C.: Case-Based Reasoning Applied to Medical Diagnosis and Treatment. In: Omatu, S., Neves, J., Rodriguez, J.M.C., Paz Santana, J.F., Gonzalez, S.R. (eds.) Distrib. Computing & Artificial Intelligence. AISC, vol. 217, pp. 137–146. Springer, Heidelberg (2013)
3. Andritsos, P., Jurisica, I., Glasgow, J.: Case-Based Reasoning for Biomedical Informatics and Medicine. In: Springer Handbook of Bio-/Neuroinformatics, Part C, pp. 207–221. Springer, Heidelberg (2014)
4. Abidi, S., Manickam, S.: Leveraging XML-based electronic medical records to extract experiential clinical knowledge an automated approach to generate cases for medical case-based reasoning systems. International Journal of Medical Informatics 68, 187–203 (2002)

5. Lee, D., Cornet, R., Lau, F., Keizer, N.: A survey of SNOMED CT implementations. Journal of Biomedical Informatics 46, 87–96 (2013)
6. Burnum, J.: The misinformation era: the fall of the medical record. Ann. Intern. Med. 110, 482–484 (1989)
7. Weiner, M., Embi, P.: Toward reuse of clinical data for research and quality improvement: the end of the beginning? Ann. Intern. Med. 151, 359–360 (2009)
8. Lei, J.: Use and abuse of computer-stored medical records. Methods Inf. Med. 30, 79–80 (1991)
9. Borges, K., Aquino, R., Barcelos, T., Simoes, J.: A methodology for preprocessing data for application of case based reasoning. In: 2012 XXXVIII Conferencia Latinoamericana En IEEE Informatica (CLEI), pp. 1–8 (2012)
10. Jayalskshmi, T., Santhakumaran, A.: Impact of Preprocessing for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks. In: IEEE Second International Conference on Machine Learning and Computing, pp. 109–112 (2010)
11. Begum, S., Ahmed, M., Funk, P., Xiong, N., Folke, M.: Case-Based Reasoning Systems in the Health Sciences: A Survey of Recent Trends and Developments. IEEE Transactions on Systems, Man, and Cybernetics, Part C 7(1), 39–59 (2010)
12. Wu, D., Weber, R., Abramson, D.: A case-based framework for leveraging nutrigenomics knowledge and personalized nutrition counseling. In: Proceeding of Workshop CBR Health Sci.s, pp. 71–80 (2004)
13. Jagannathan, R., Petrovic, S.: Dealing with Missing Values in a Clinical Case-Based Reasoning System. In: Second IEEE International Conference on Computer Science and Information Technology (ICCSIT), pp. 120–124 (2009)
14. Floyd, M.W., Davoust, A., Esfandiari, B.: Considerations for Real-Time Spatially-Aware Case-Based Reasoning: A Case Study in Robotic Soccer Imitation. In: Althoff, K.-D., Bergmann, R., Minor, M., Hanft, A. (eds.) ECCBR 2008. LNCS (LNAI), vol. 5239, pp. 195–209. Springer, Heidelberg (2008)
15. Weiskopf, N., Weng, C.: Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal American Medical Informatics Association 20(1), 144–151 (2013)
16. Klompas, M., Eggleston, E., McVetta, J., Lazarus, R., Li, L., Platt, R.: Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. Diabetes Care 36(4), 914–921 (2013)
17. Esfandiari, N., Babavalian, M., Moghadam, A., Tabar, V.: Knowledge discovery in medicine: Current issue and future trend. Expert Systems with Applications 41, 4434–4463 (2014)
18. Kuhn, M., Johnson, K.: Data Pre-processing. Applied Predictive Modeling, 27–59 (2013)
19. Xie, X., Lin, L., Zhong, S.: Handling missing values and unmatched features in a CBR system for hydro-generator design. Computer-Aided Design 45, 963–976 (2013)
20. Guessouma, S., Laskrib, M., Lieberc, J.: RespiDiag: A Case-Based Reasoning System for the Diagnosis of Chronic Obstructive Pulmonary Disease. Expert Systems with Applications 41(2), 267–273 (2014)
21. Piramuthu, S.: Evaluating feature selection methods for learning in data mining applications. European Journal of Operational Research 156(2), 483–494 (2004)
22. Baig, M.: Case-based reasoning – an effective paradigm for providing diagnostic support for stroke patients. Master Thesis, Queen's University, Kingston, Ontario, Canada (2008)
23. Bottrighi, A., Leonardi, G., Montani, S., Portinale, L., Terenziani, P.: Intelligent Data Interpretation and Case Base Exploration through Temporal Abstractions. In: Bichindaritz, I., Montani, S. (eds.) ICCBR 2010. LNCS (LNAI), vol. 6176, pp. 36–50. Springer, Heidelberg (2010)

24. Gopal, K.: Efficient case-based reasoning through feature weighting, and its application in protein crystallography. Ph.D. Thesis,Texas A & M University (2007)
25. Xiong, N., Funk, P.: Combined feature selection and similarity modelling in case-based reasoning using hierarchical memetic algorithm. In: IEEE Congress on Evolutionary Computation (CEC), pp. 1–6 (2010)
26. Shanga, C., Min, M., Fenga, S., Jianga, Q., Fana, J.: Feature selection via maximizing global in-formation gain for text classification. Knowledge-Based Systems 54, 298–309 (2013)
27. Huang, Y., McCullagh, P., Black, N., Harper, R.: Feature selection and classification model construction on type 2 diabetic patients' data. Artificial Intelligence in Medicine 41, 251–262 (2007)
28. Kwiatkowska, M., Atkins, S.: Case Representation and Retrieval in the Diagnosis and Treatment of Obstructive Sleep Apnea: A Semio-fuzzy Approach. In: Proceedings of 7th European Case Based Reasoning Conference (ECCBR), pp. 25–35 (2004)
29. Balakrishnan, V., Shakouri, M., Hoodeh, H.: Integrating association rules and case-based reason-ing to predict retinopathy. Maejo International Journal of Science and Technology 6(03), 334–343 (2012)
30. Kotsiantis, S., Kanellopoulos, D., Pintelas, P.: Data Preprocessing for Supervised Leaning. International Journal of Computer Science 1(2), 111–117 (2006)
31. Han, J., Rodriguze, J., Beheshti, M.: Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner. In: 2008 Second International Conference on Future Generation Communication and Networking, pp. 96–99 (2008)
32. Pla, A., López, B., Gay, P., Carles, C.: eXiT*CBR.v2: Distributed case-based reasoning tool for medical prognosis. Decision Support Systems 54(3), 1499–1510 (2013)
33. Rea, S., Pathak, J., et al.: Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. Journal of Biomedical Informatics 45, 763–771 (2012)
34. McSherry, D.: Precision and Recall in Interactive Case-Based Reasoning. In: Aha, D.W., Watson, I. (eds.) ICCBR 2001. LNCS (LNAI), vol. 2080, pp. 392–406. Springer, Heidelberg (2001)
35. Wang, X., Dong, J.: Fuzzy Based Similarity Adjustment of Case Retrieval Process in CBR System for BOF Oxygen Volume Control. In: Sixth International Conference on Advanced Computational Intelligence, Hangzhou, China, pp. 19–21 (2013)
36. The myCBR3 Project, `http://www.mycbr-project.net/download.html` (last accessed on May 11, 2014)
37. RapidMiner, `http://rapidminer.com/` (last accessed on May 20, 2014)
38. HL7 Version 3: Reference Information Model (RIM), `http://www.hl7.org` (last accessed on May 11, 2014)
39. Molina, L.C., Belanche, L., Nebot, A.: Feature selection algorithms: a survey and experimental evaluation. In: IEEE International Conference on Data Mining ICDM, pp. 306–313 (2002)
40. Kar, D., Chakraborti, S., Ravindran, B.: Feature Weighting and Confidence Based Prediction for Case Based Reasoning Systems. In: Agudo, B.D., Watson, I. (eds.) ICCBR 2012. LNCS, vol. 7466, pp. 211–225. Springer, Heidelberg (2012)
41. Jagannathan, R., Petrovic, S.: Dealing with missing values in a clinical case-based reasoning system. In: Second IEEE International Conference on Computer Science and Information Technology (ICCSIT), pp. 120–124 (2009)
42. Michael, F.: Microvascular and Macrovascular Complications of Diabetes. American Diabetes Association, Clinical Diabetes 26(2), 77–82 (2008)
43. Hassanien, A.E., Abdelhafez, M.E., Own, H.S.: Rough sets data analysis in knowledge discovery: a case of kuwaiti diabetic children patients. Advances in Fuzzy Systemsol. 8, 2 (2008)