2011 IEEE International Conference on Fuzzy Systems
June 27-30, 2011, Taipei, Taiwan

Improving Disease Prediction Using ICD-9 Ontological Features

Mihail Popescu
Health Management and Informatics Department
University of Missouri
Columbia, MO 65211 USA
popescum@missouri.edu

Mohammad Khalilia
Computer Science Department
University of Missouri
Columbia, MO 65211 USA
mohammedsk@gmail.com

Abstract— **Disease prediction has become important in a variety of applications such as health insurance, tailored health communication and public health. Disease prediction is usually performed using publically available datasets such as HCUP, NHANES or MDS that were initially designed for health reporting or health cost evaluation but not for disease prediction. In these datasets, medical diagnoses are traditionally arranged in "diagnose-related groups" (DRGs). In this paper we compare the disease prediction based on crisp DRG features with the results obtained employing a new set of features that consist of the fuzzy membership of patient diagnoses in the DRG groups. The fuzzy membership features were computed using an ICD-9 ontological similarity approach. The prediction results obtained on a subset of 9,000 patients from the 2005 HCUP data representing three diseases (diabetes, atherosclerosis and hypertension) using two classifiers (random forest and SVM trained on 21,000 samples) show significant (about 10%) improvement as measured by the area under the ROC curve (AROC).**

Keywords-component; disease prediction; ICD-9 similarity measure; ontological features; random forest; SVM.

I. INTRODUCTION

Disease prediction is employed in different domains such as risk management, tailored health communication and public health. Risk management plays an important role in health insurance industry, mainly in the underwriting process. Health insurers use a process called underwriting in order to classify the applicant as standard or substandard, based on which they compute the policy rate and the premiums individuals have to pay [1].

Another domain where disease prediction may be applied is tailored health communication. For example, we can target specific medical educational materials and news to a subgroup within the general population that has a high predicted risk for a given disease. Cohen et al [2] discussed how tailored health communication for cancer patients can motivate cancer prevention and early detection. Disease risk prediction along with tailored health communication represents an effective preventive medicine method that may lead in the long-run to a reduction in the cost of medical care.

The reporting requirements of various US governmental agencies such as Center for Disease Control (CDC), Agency for Health Care Quality (AHRQ) and US Department of Health and Human Services Center for Medicare Services (CMS) have created huge public datasets that, we believe, are not utilized to their full potential. For example, CDC (www.cdc.gov) makes available National Health and Nutrition Examination Survey (NHANES) data. Using NHANES data, Yu et al. [3] predicts diabetes risk using an SVM classifier. CMS (www.cms.gov) uses the Medicare and Medicaid claims to create the minimum dataset (MDS). Herbert et al. [4] uses MDS data to identify people with diabetes. In this paper we use the National Inpatient Sample (NIS) data created by AHRQ (www.ahrq.gov) Healthcare Utilization Project (HCUP), to predict the risk for three diseases: diabetes, atherosclerosis and hypertension. To compute the disease risk we use a new set of ICD-9 features based on ontological similarity between the ICD-9 diagnoses contained in a DRG and the ICD-9 diagnoses of the patient. We compare this approach with the prediction of the same diseases described in [5].

The remainder of this article is structured thusly: in Section II we describe the ICD-9 medical taxonomy together with the similarity measure used in the feature extraction process. In Section III we describe the proposed ontological feature extraction algorithm together with a brief description of the classifiers employed, in Section IV we show some results obtained on a subset of 2005 HCUP patient dataset and in Section V we provide some conclusions and ideas for future research.

II. ICD-9 MEDICAL TAXONOMY AND SIMILARITY MEASURE

International classification of diseases-version 9 (ICD-9) is a diagnose coding system [6] used in hospitals for data retrieval and billing purposes. Every code represents a disease, condition, symptom, or cause of death. However, from our point of view ICD-9 represents an ontology, i.e. a controlled vocabulary overlaid with a "is-a" term hierarchy. The controlled vocabulary allows for detection of synonymy when two diagnoses are compared. The hierarchy (tree) structure allows for assessing the semantic similarity between diagnoses. A snippet of the ICD-9 tree is shown in Fig. 1.

This work has been supported in part by a UM Research Board grant.

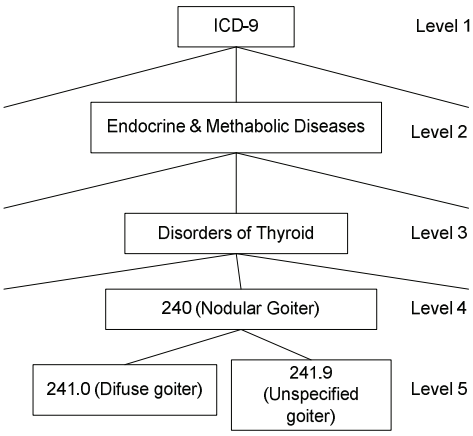

Figure 1. Partial view of the ICD-9 hierarchy

From Fig. 1 we see that diagnoses 241.0 and 241.9 are semantically related although they are syntactically (string-wise) different. Unfortunately, the ICD-9 hierarchy has only 5 levels. This will have an impact on the granularity of the term similarity as many pairs of terms will have the same similarity coefficient. However, even with this low granularity the impact on prediction performance is significant.

In our view, the hierarchical structure of the ICD-9 ontology represents the knowledge of the medical field as viewed by the domain experts (physicians). The key of our approach is to use the domain knowledge (hierarchy) in computing patient similarities. Given two patients described by a sequence of ICD-9 diagnoses, T_{ij}, we first consider the problem of computing the association (seen as fuzzy membership), $s(T_1, T_2)$, between two terms (diagnoses) T_1 and T_2.

There are many algorithms for defining term similarity in a taxonomy (see Chapter 2 in [7]). One way of computing term similarity is to assign each term T_i weights based on its importance, IMP, within the ontology. As a consequence, two patients are more similar if they both have the same rare (in the database) disease (say cystic fibrosis) than if they both have flu. The term importance can be computed (see Chapter 2 in [7]) using path-based, depth-based, density-based, information content-based approaches. In this work we use a depth-based approach. The importance IMP, of a term in the ICD-9 taxonomy is computed as IC=1-1/n where n = {1, 2, 3, 4, 5} is the level of the term within the hierarchy. For example, the IMP of diagnosis code 241 (Goiter, level 4) is IMP(241) = 1-1/n = 1-1/4 = 0.75. For consistency, we consider IMP(level 5) = 1. Now, returning to the problem from the beginning of this paragraph, the similarity of two diagnosis terms, $s(T_1, T_2)$, is defined as:

$$s(T_1, T_2)=IMP(NCA(T_1, T_2)). \quad (1)$$

where NCA="nearest common ancestor" of the two terms in the ontology.

In Fig. 1, the nearest common ancestor (NCA) of 241.0 and 241.9 is 241, and so, the similarity between the two diagnoses is s(241.0, 214.9)=IMP(241)=0.75. This is clearly the simplest approach and it is only used here for illustrative purposes.

For two sets of ICD-9 terms, $P_1 = \{T_{11}, \ldots, T_{1n}\}$ and $P_2=\{T_{21}, \ldots, T_{2m}\}$ we can define a variety of similarities (see Chapter 2 in [7] for details). In this paper we consider the following simple formula:

$$s(P_1, P_2)=\max_{ij}\{s(T_{1i}, T_{2j})\}. \quad (2)$$

III. STUDY METHODOLOGY

To better understand the feature extraction process we first describe the 2005 HCUP dataset used in this paper (denoted henceforth HCUP2005).

A. The HCUP2005 dataset

The Nationwide Inpatient Sample (NIS) is a database of hospital inpatient admissions that dates back to 1988 and it is used to identify, track, and analyze national trends in health care utilization, access, charges, quality, and outcomes. The NIS database is developed by the Healthcare Utilization Project (HCUP) and sponsored by the AHRQ. This database is publicly available and does not contain any patient identifiers. The database contains discharge level information on all inpatients from a 20% stratified sample of hospitals across the United States, representing approximately 90% of all US hospitals [8]. HCUP data from the year 2005, denoted as HCUP2005, will be used in this paper.

The data set contains 7,995,048 hospital stays and 126 clinical and nonclinical data elements for each hospital stay. Nonclinical elements include patient demographics, hospital identification, admission date, zip code, calendar year, total charges and length of stay. Clinical elements include procedures, procedure categories, diagnosis codes and diagnosis categories. Every record contains a vector of 15 ICD-9 diagnosis codes. In addition, every record contains a vector of 15 diagnosis category codes (DRGs). The diagnosis categorization is performed using the Clinical Classification Software (CCS) developed by HCUP. There are numerous ICD-9 codes, over 14,000, in HCUP; CCS collapsed these codes into a smaller number of clinically meaningful DRGs. There are 259 diagnosis categories in the HCUP2005 dataset, every category is denoted by a value in the range 1-259. Demographics such as age, race and sex are also included in our data set and used predicting the three medical conditions.

The prevalence in the HCUP2005 dataset of three diseases used in this paper (hypertension, diabetes mellitus and coronary atherosclerosis) is given in table I below. As we see from table I, some diseases like testis cancer might not have a sufficient number of samples for training a classifier even on such a large dataset.

TABLE I. THE PREVALENCE OF FOUR DISEASES IN THE HCUP2005 DATASET

Disease	Prevalence
Hypertension	29.1%
Diabetes mellitus, no complications	12%
Coronary Atherosclerosis	27.65%
Testis Cancer	0.046%

B. ICD-9 based Ontological Features

To predict a disease, we extract from HCUP2005 a random set of N patients, N/2 with the disease and N/2 without it. For each patient, P_i with i = {1, … , N}, we used from HCUP2005 the following variables: age, race, sex, 15 ICD-9 codes ($P_{i,ICD9}$) and 15 diagnosis categories ($P_{i,DRG}$). As mentioned before, there are 259 DRGs, DRG_j with j = {1, …, 259}, and every group contains a set of ICD-9 codes, DRG_j = {$ICD9_{j1}$, …, $ICD9_{jk}$}. In [5] we represented each patient P_i using a feature vector $\mathbf{x}_i^{crisp} \in R^P$, with P=262 dimensions. Features 1-259 (DRG related) were computed as:

$$x_{ip}^{crisp} = \begin{cases} 1 & P_{i,ICD9} \cap DRG_p \neq \varnothing \\ 0 & P_{i,ICD9} \cap DRG_p = \varnothing \end{cases} \quad (3)$$

Essentially, since the DRGs were computed for us, the feature vector had an 1 in position p if DRG_p was contained in the diagnoses set of patient i, $P_{i,DRG}$. As a result, each feature vector contained at most 15 ones (the number of DRGs stored per patient) which was a rather sparse representation. The last 3 features (index 260, 261, 262) were sex, age and race, respectively.

In this paper we propose to compute the 259 diagnose related features using a fuzzy membership in each DRG_j, i.e. a value between 0 and 1 that represents the similarity between a diagnose and DRG_j. The proposed features will be calculated as:

$$x_{ip} = s(P_{i,ICD9}, DRG_p), p \in \{1, \ldots, 259\}, \quad (4)$$

where the above similarity, s, is computed using formula (2). Features with index 260, 261 and 262 are similar to the ones we used in [5], i.e. age, race and sex.

Example 1

Consider a patient with the following set of diagnoses P={682.6, 486, 453.42, 493.92,41.11,250.00,782.3,278.01}. The crisp [5] and the fuzzy (ontological) features (index 1-259) related to this patient are shown in Fig. 2.

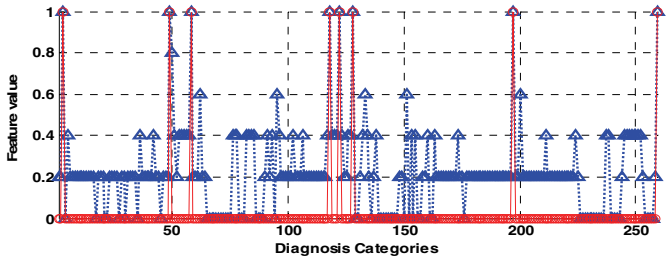

Figure 2. The crisp (red) and fuzzy ontological (blue) features for patient P from example 1.

We note that the ontological features coincide with the crisp ones for DRGs that contain one of the ICD-9 codes from the diagnoses set (in number of 8). An example is DRG index 49 ("diabetes mellitus without complications") that contains the ICD-9 code 250.00 ("Diabetes mellitus without complication type II or unspecified type not stated as uncontrolled"). However, there are other indices where the ontological features have a high value. Take for example DRG index 50 ("diabetes mellitus with complications") that contains among others the ICD-9 code 250.03 ("Diabetes mellitus without complication type I uncontrolled"). Since P does not contain code 250.0, $x_{50}^{crisp} = 0$. However, since s(250.00, 250.03) = 0.8, the related ontological feature is greater than zero, i.e. $x_{50} = 0.8$. Aside from the fact that by using the relations from the ICD-9 taxonomy we provide a better representation of the diagnoses set, we also account for the uncertainty of the coding process itself (known to be somewhat unreliable).

C. Classifiers used

In this paper we present experiments performed with two classifiers, random forests (RF) [9] and support vector machines (SVM) [10].

RF is an ensemble learner, a method that generates many classifiers and aggregates their results. RF adds a layer of randomness to bagging by building large collection of de-correlated trees. RF will create multiple CART-like trees, each trained on a bootstrap sample of the original training data and searches across a randomly selected subset of input variables to determine the split. Each tree in RF will cast a vote for some input x, then the output of the classifier is determined by majority voting of the trees. Since the focus of this paper is on features rather than on classifiers themselves, we refer the reader to [9] and [10] for more details on RF and SVM.

We performed the classification using R, which is an open source statistical software. We used R *randomForest* and *SVM* (e1071) packages. The parameters to the RF were as follows: number of trees (ntree) was set to 500. Overall, the number of trees didn't seem to influence the classification results. The number of variables randomly sampled as candidates at each split (mtry) is equal to the square root of the number of features. Since in our case we have 262 features, mtry was consequently set to 16.

For SVM we used a linear kernel, termination criterion (tolerance) was set to 0.001, epsilon for the insensitive-loss function was 0.1 and the regularization term (cost) was 1.

D. Experiments

We tested both classifiers, RF and SVM, with N=10,000 patients extracted from HCUP2005, N_d=5,000 that had the disease and N_n=5,000 that didn't. Out of the 10,000 samples, 7,000 were used for training and 3,000 for testing. Obviously, we excluded the target disease from the diagnosis set of the N_d patients that had it. We performed the same experiment for three diseases (first 3 lines in table I): hypertension, diabetes mellitus and arteriosclerosis. The results obtained are showed in the next section.

IV. Results

The classification results, area under the curve (AROC) and receiver operating characteristic (ROC) curve, for diabetes are shown in table II and Fig. 3.

TABLE II. AROC RESULTS FOR DIABETES PREDICTION

	Crisp Features	Fuzzy Features
RF	0.9524	0.9996
SVM	0.8567	0.981

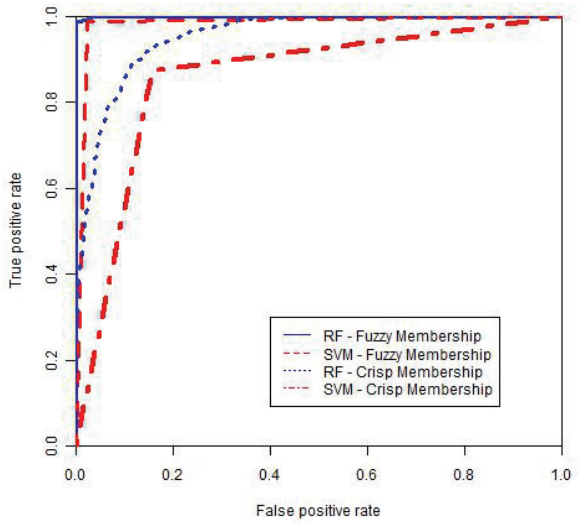

Figure 3. ROC curves for diabetes prediction obtained with random forest (blue) and SVM (red).

The AROC improvement was smaller for RF than for SVM, since RF had already a good prediction performance dues to its builtin feature selection property.

The results obtained for atheoriosclerosis prediction are shown in table III and Fig. 4.

TABLE III. AROC RESULTS FOR ARTHERIOSCLEROSIS PREDICTION

	Crisp Features	Fuzzy Features
RF	0.9647	0.9995
SVM	0.8833	0.9737

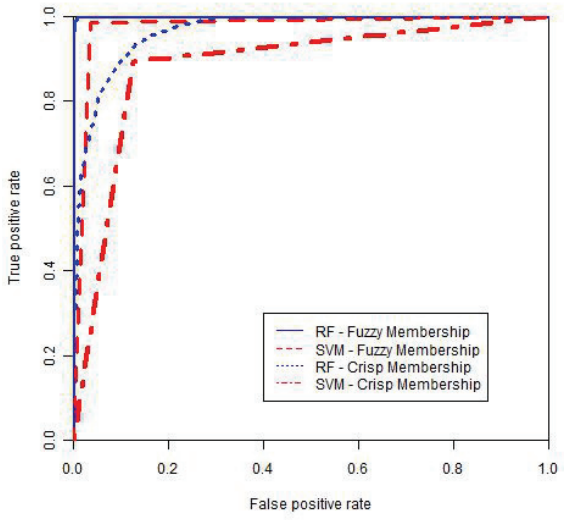

Figure 4. ROC curves for aterioscleroses prediction obtained using random forest (blue) and SVM (red)

For the atherioscelrosis prediction, too, we obtained a significant performance improvement (3-9%) when fuzzy features are used.

The results obtained for hypertension prediction are shown in table IV and Fig. 5. Again, a notable AROC improvement (5-13%) is obtained by using the fuzzy features instead of the crisp ones.

TABLE IV. AROC RESULTS FOR HYPERTENSION PREDICTION

	Crisp Features	Fuzzy Features
RF	0.9454	0.9991
SVM	0.8537	0.989

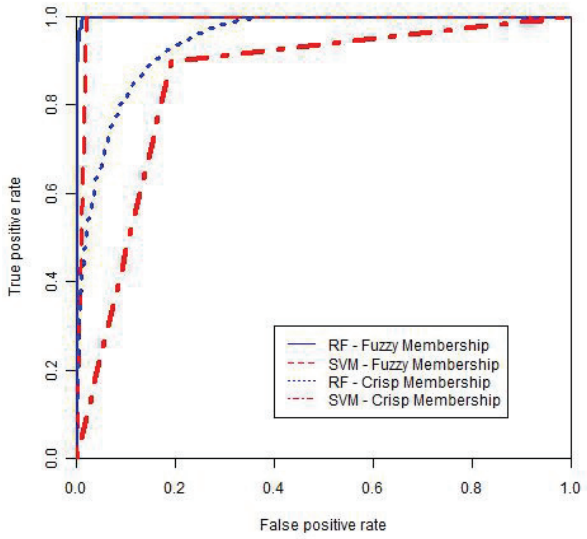

Figure 5. ROC curves for hypertension prediction obtained using random forest (blue) and SVM (red)

V. CONCLUSIONS

In this paper we presented a method for disease predicting using large public medical datasets. Disease prediction is important in a variety of applications such as health insurance, tailored health communication and public health. The presented method is based on employing ICD-9 diagnostic groups (DRGs) and demographics variables in conjunction with classification algorithms, such as SVM and RF. As opposed to using a crisp DRG membership for the ICD-9 codes, we introduced a novel fuzzy membership computed based on ICD-9 ontological similarity. The results presented on three different diseases and two classifiers show that the fuzzy features lead to an important improvement (between 3 and 13%) in prediction performance. The improvement is due to the fact that the fuzzy features capture the relationships between the DRG groups in the process of feature extraction.

In the near future, we plan to extend our approach to all 259 disease categories present in the HCUP dataset.

REFERENCES

[1] T. Yi, Z. Guo-Ji. The application of machine learning algorithm in underwriting process", Proceedings of the 4th International Conference on Machine Learning and Cybernetics, Guangzhou, China, Aug. 2005, p. 3524-3527.

[2] E. Cohen, C. A. Caburnay, D.A. Luke, S. Rodgers, G.T. Cameron, M.W. Kreuter, "Cancer coverage in general-audience and black newspapers", *Health Communication*, 2008. 23(5): p. 427-435.

[3] W. Yu, T. Liu, R. Valdez, M. Gwinn, M.J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes", *BMC Medical Informatics and Decision Making*, 2010. 10(1): p. 16.

[4] P. Hebert, L. Geiss, E. Tierny, M. Engelgau, B. Yawn, A. McNeab, "Identifying persons with diabetes using Medicare claims data", *American Journal of Medical Quality*, 1999. 14(6): p. 270.

[5] M. Khalilia, M. Popescu, "Predicting disease risks from highly unbalanced data using random forest", submitted to BMC Medical Informatics and Decision Making, 2011.

[6] http://www.cms.hhs.gov/ICD9ProviderDiagnosticCodes

[7] M. Popescu, D. Xu, Eds., Data Mining in Biomedicine using ontologies, Artech House, Norwood, MA 2009.

[8] HCUP (2009), Overview of the Nationwide Inpatient Sample (NIS), www.ahrq.com.

[9] L. Breiman, "Random Forests", *Machine Learning*, 45 (1):5-32.

[10] C. Cortes and V. Vapnik, "Support-Vector Networks", *Machine Learning*, 20, 1995.