Domain Ontology-based Feature Reduction for High Dimensional Drug Data and its Application to 30-Day Heart Failure Readmission Prediction

(Invited Paper)

Sisi Lu[1, 2], Ye Ye[1], Rich Tsui[1], Howard Su[1], Ruhsary Rexit[1, 2], Sahawut Wesaratchakit[1], Xiaochu Liu[3], Rebecca Hwa[2]
[1] The RODS Laboratory, Department of Biomedical Informatics [2] Department of Computer Science
University of Pittsburgh, Pittsburgh, PA 15260
[3] Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093
{sil21, yey5, tsui2, hos10, saw111}@pitt.edu {ruhsary, hwa}@cs.pitt.edu xil174@cs.ucsd.edu

***Abstract*—High dimensional feature space could potentially hinder the efficiency and performance for machine learning, and high correlations between features may further increase the redundancy and diminish performance of learning algorithms. Domain ontology provides relationships and similarities between concepts in the specific area, and thus can be used to reduce redundancy by clustering concepts and revealing their functionality. In this paper, we study the problem of using high dimensional medication data to predict the probability of 30-Day heart failure readmission. We propose a feature reduction method for high dimensional dataset using a combination of two drug ontologies. By creating a tree structure of the combination, the method uses a greedy strategy to obtain a subset of features, which may have higher correlation with the class label but lower correlation with each other. Experimental results show that our methods improve the performance of heart failure readmission prediction (using only drug data) comparing to existing feature reduction methods without drug domain ontologies.**

Keywords: High Dimensional Data, Feature Reduction,Feature Selection, Domain Ontology, Heart Failure Readmission Prediction.

I. Introduction

Heart failure readmission has imposed a large financial burden to healthcare in the United States. According to published data from Agency for Healthcare Research and Quality (AHRQ) for fiscal year 2010, the national rate of readmission for heart failure patients reached 25% with the mean cost per readmission to be $13,680 [1]. It is profitable for hospital readmission reduction teams and insurance providers to identify high-risk readmission group through analysis of routinely collected electronic health records (EHRs) when patients stay in the hospital.

One type of routinely collected EHRs is the in-patient prescription medication, which contains a significant amount of clinical information about the patient's health status. We believe it can be used to train a classifier to predict whether a patient with a heart failure might be readmitted within 30 days. In this pilot, we focus on using only the prescription drug list because the dataset we collected is already very large in size and in dimension such that it is difficult to process them directly. The goal of this work is to reduce the dimensionality of the dataset so that it is practically feasible to train a classifier to predict 30-day readmissions. In the future, we may incorporate additional types of information from the EHR.

One of the challenges for working with "big data" is that the features in the dataset may have a high dimensionality and be sparsely represented. This could potentially hinder the efficiency and performance for machine learning. In our problem, for example, the University of Pittsburgh Medical Center Health System (UPMC-HS) employs 11,253 unique drug codes, and each patient could have up to 40 different drug codes administered during his/her hospital stay. Thus, the dataset built directly from these drug codes will have a high dimensionality and be very sparse. High dimensional data usually include a large number of irrelevant, redundant or abnormal information, which can dramatically diminish the performance of learning algorithms [2]. These impurity and noise in the dataset would damage the performance of machine learning algorithms.

The main contribution of this work is a method for reducing the dimensionality of the prescription medication dataset through the use of drug ontologies as domain knowledge resources. Ontological relationships can help to cluster the drugs by a set of criteria including therapeutic intention, active ingredient, precise clinical name and others. Our approach is a three-step process. First, we combine two publicly accessible drug ontologies (RxNorm and NDF-RT) into a coherent tree-structured hierarchy. Then, we apply a greedy-based search strategy to select a subset of nodes in the tree-structured drug ontology as filtered features. Finally, we apply the information gain ratio as a measurement to filter the features.

This paper is organized as follows. Section II lists related work on feature selection using domain ontology and heart failure readmission prediction problem. Section III presents the method to build tree-structure drug ontology, feature dimensional reduction method and evaluation method. Section IV describes the experiment setup and results. Section V concludes the paper with future work.

COLLABORATECOM 2013, October 20-23, Austin, United States

DOI 10.4108/icst.collaboratecom.2013.254124

II. Related Work

Heart failure readmission prediction is still a challenge. Current literatures show the area under the ROC curve for 30-day readmission ranges between 0.60 and 0.69 [3]. High performance models may rely on many features that are not routinely collected, (e.g. number of home address changes) [4] or some features that are not easy to be automatically retrieved (e.g. history of percutaneous coronary intervention) [5]. Therefore, those models can have limited use for prediction systems using routinely collected data. We use routinely collected prescription medication during patients stay from Medical Archive System (MARS) hospital pharmacy data repositories. Then, we employ drug ontologies to reduce the feature dimension to a feasible size. The whole procedure can be done in online mode. Therefore, it can be integrated into automatically readmission prediction systems and could be helpful to reduce the readmission possibility.

Dimension reduction algorithms can fall into two categories: feature selection and feature extraction [6]. Feature selection is the process of identifying a subset of features in the train dataset that might be most helpful for machine learning algorithms. Generally, there are two types of feature selection methods: wrapper model and filter model. The wrapper model is to use a particular classifier (learning algorithm) as evaluator to get a subset features that most suitable for that classifier. The filter model is to select a subset of features without the dependency on certain classifier. For example, information gain (IG), information gain ratio (GR), χ^2 statistics can be used to as measurement for filter models [7]. Feature extraction transforms the original dataset into a smaller set of new features that still capture most of the information [2]. For example, principal components analysis (PCA) can be used as feature extraction model, which uses the dependencies between variables to represent the high-dimensional data in a more tractable, lower-dimensional form, without losing much information. Domain knowledge resources are not included in these dimension reduction algorithms. Our drug ontology based feature dimension reduction method employs drug ontology as knowledge resource combining the correlation measurement heuristics, which is more feasible to our drug dataset.

Domain ontology guided feature-selection method has been proposed to do document categorization. In [8], researchers mapped original terms to concepts based on Unified Medical Language System (UMLS) knowledge, used a bottom-up hill climbing search algorithm to find an optimal subset of concepts, and showed improved accuracy of a KNN classifier. However, the bottom-up feature search method uses heuristic function to test whether includes a feature or not. The heuristic function uses a KNN classifier, which calculates the similarity between each pair of instances in the dataset. It is not feasible for our dataset with large number of instance and feature as its time complexity.

III. Methods

In this section, we first present the hierarchical drug ontology that we constructed by combining two publicly available drug ontologies: namely the RxNorm and the National Drug File Reference Terminology (NDF-RT). Then, we demonstrate a greedy based top-down search strategy to do feature selection based on the hierarchical drug ontology we constructed. Finally, we introduce the learning model used for evaluation and evaluation criterion.

A. Constructing a Hierarchical drug ontology using RxNorm and NDF-RT

Drug ontologies are important to communicate, compare and sharing drug data among different systems. Moreover, they can provide semantic meanings, functionality, and relationship of drugs [8]. However, because of the inherent complexity and broad scope of drug domain knowledge, the coverage of single drug ontology can be limited. Also, due to the incomplete coverage issues, the combination among different drug ontologies is not a trivial task. In this work, we assume that therapeutic intention of drugs, general active ingredients, dosage, and dose form may be important factors to the heart failure readmission prediction. Thus, we utilize RxNorm and NDF-RT to get a new drug ontology, where RxNorm serves as drug thesaurus and NDF-RT serves as drug functionality knowledge base.

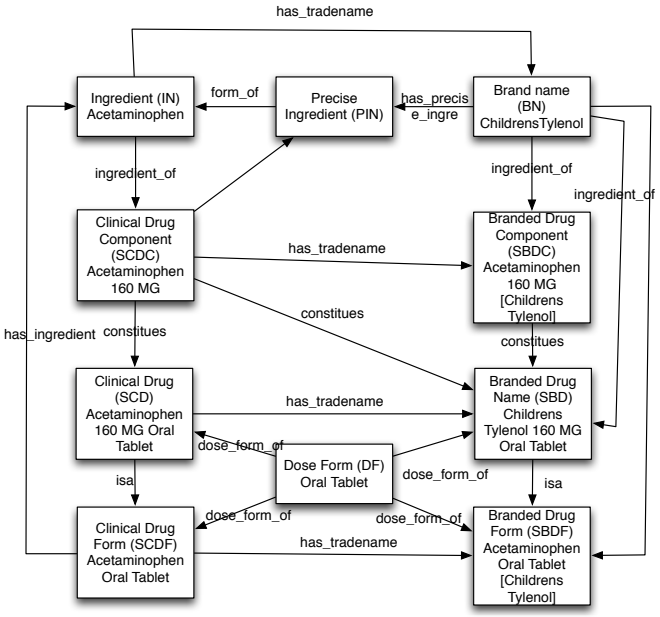

Figure 1 RxNorm hierarchy organization (adopted from [13])

RxNorm is a standardized drug ontology and terminology system maintained by the US National Library of Medicine [9]. It provides normalized naming system for both generic and branded drugs [10]. As it defines unique identifier for medicines and drugs from a various number of sources, it can be used to exchange and communicate between different drug terminology systems efficiently. Besides, RxNorm provides relationships between generic drugs, branded drugs, clinical drug component, precise ingredient, generic name ingredient and so forth. However, it does not provide the interaction between drugs, or therapeutic drug classification purpose. Fig. 1 shows the hierarchy structure of RxNorm.

NDF-RT is a drug ontology from the Veterans Health Administration [11]. Like RxNorm, NDR-RT also provides

relationship between drugs and ingredients. Moreover, it provides two lists of drug classes: Legacy VA classes and External Pharmacologic classes [8]. Fig. 2 shows the hierarchy structure of NDF-RT. Legacy VA classes can be used as clinically oriented drug classification, which are organized into a shallow hierarchy [12] from one level up to three levels. Specifically, Legacy VA Classes classify drugs according to chemical or pharmacological function, or by therapeutic function. A clinical drug (for VA Products level) is assigned to only one VA class. Therefore, clinical drugs can be classified according to VA classes. However, the coverage of NDF-RT is not as good as RxNorm. Approximately 54% of RxNorm drug concepts cannot find its corresponding NDF-RT concepts [9]. Most of these miss-mappings are due to the differences in dosage, strength, route form, and brand names [13].

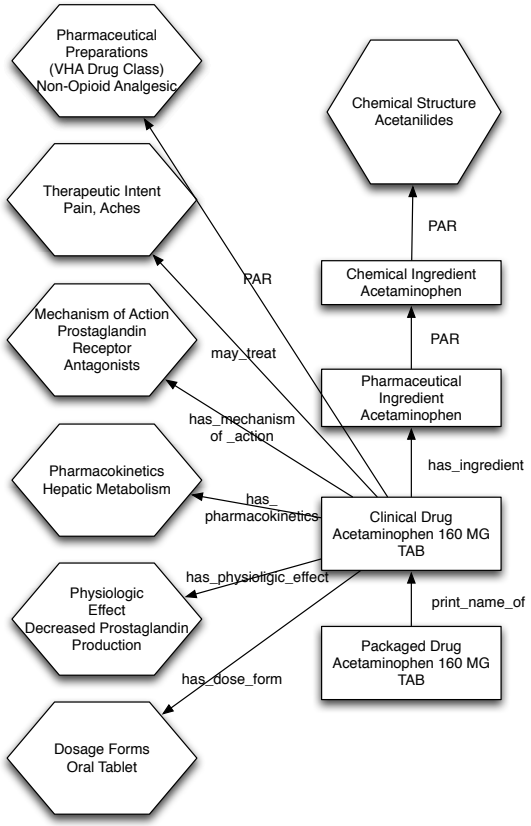

Figure 2 National Drug File-Reference Terminology drug-class hierarchy organization (adapted from [8])

One challenge of using patient prescription information is that data are all coded with local coding systems of each hospital and the hierarchical relationships of these non-standardized codes are not available in any knowledge resources (e.g. RxNorm and NDF-RT).

To solve this problem, we first use the code description information (e.g. clinical drug name) to find the most approximate RxNorm concept unique identifier (RXCUI). If no RXCUI can be found using the full clinical drug names, we truncate the dosage part of the description to find the corresponding RXCUI. For example, if we cannot find the RXCUI for drug "*GATIFLOXACIN 0.5% OPH SOLN 2.5ML*", we will use "*GATIFLOXACIN 0.5% OPH SOLN*" to search for its corresponding RXCUI.

Then, we employ a two-stage approach proposed in [13] to get approximately 93% of RXCUI mapping to NDF-RT unique identifier (NUI). Specifically, we first find the ingredient given a RXCUI code by searching the relationship defined in RxNorm. Then, we use the RXCUI of the ingredient to find its corresponding NUI. For the mapped NUI, we find the VA Class using the relationship between ingredients and clinical drugs in NDF-RT.

We create a tree structure to represent the hierarchical relationship between drugs based on ontology knowledge. This tree structure is derived from a combination of RxNorm and NDF-RT. RxNorm uses five levels for generic drugs: Ingredient (IN), Precise Ingredient (PIN), Clinical Drug Component (SCDC), Clinical Drug From (SCDF) and Clinical Drug (SCD). The differences between these five layers are as respect to the active ingredients, strength, and dose form [13]. Branded name (BN), Branded Drug Component (SBDC), Branded Drug Form (SBDF), Branded Name with strength and dose form (SBD) are the corresponding levels for branded name drug concepts. We assume that this graph structure is overcomplicated for modeling drug information in the heart failure readmission prediction problem. To reduce the complexity, we construct a new hierarchy structure combining RxNorm and NDF-RT Legacy VA Classes, as shown in Fig. 3. We add a pseudo node as the root node. The top three levels are NDF-RT Legacy VA Classes, as there are at most three levels in legacy VA Classes. For a particular clinical drug, there exits at most one corresponding legacy VA Class. These three levels represent the therapeutic intention of drugs. The fourth level represents the ingredient or ingredient groups of drugs. For a certain clinical drug, only one active ingredient or ingredient group is mapped to that drug. This level refers to the general active ingredients of drugs. The fifth level represents the RXCUI of drugs. There exists only one RXCUI given a certain clinical drug. This level refers the dosage of drugs. The lowest level represents the local drug code used by different hospitals. This level refers the dose form of drugs. According to the many-to-one mapping, this hierarchy ontology has a tree structure.

B. Greedy based top-down search strategy for feature seleciton

Our drug dataset is a high dimensional sparse imbalanced dataset. In order to utilize the tree structure drug ontology described in the previous section to do feature selection for the drug dataset, we propose a greedy based top-down search strategy. The strategy is shown in Fig. 4.

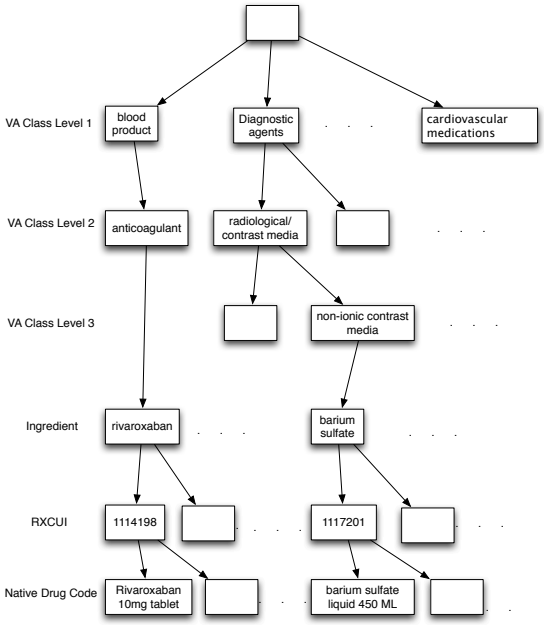

Figure 3 Drug Ontology hierarchy structure constructed by combining RxNorm and NDF-RT

The greedy based top-down search algorithm is under the assumption that a good feature subset should contain features that are highly correlated with the target variable, yet uncorrelated with other features in the subset. The input of the algorithm includes the train dataset with class labels and the tree-structure drug ontology described in the previous section. The output of the algorithm is the selected subset of features. The features in the input dataset are local drug codes, and the features in the output dataset are nodes in the tree-structure drug ontology. That is, we clustering features in the original dataset based on the drug ontology.

The algorithm will perform the following routine.

First, traverse the drug ontology tree in Depth-first search (DFS) to get each the branch of the children of the pseudo root node. Based on the construction of the drug ontology, the children of the root are the first legacy VA Classes.

Second, for each branch, we sort the nodes according to *Heuristic_Function*. The *Heuristic_Function* we used in the experiment is the *gain ratio measurement* [14]. For a feature given the dataset, the gain ration of that feature can be calculated as follows:

$$GR(F) = \frac{IM(F)}{IV(F)} \tag{1}$$

$IM(F)$ represents the information gain for Feature F, which is defined in [14]. $IV(F)$ represents the information value of Feature F, which are defined as follows:

$$IV(F) = -\sum \frac{x_i}{N} \log\left(\frac{x_i}{N}\right) \tag{2}$$

where x_i is the i th value of feature F; N is the size of the dataset. The Information gain helps to distinguish features which are most helpful to discriminating the class labels in a given train dataset. However, it prefers features with multiple distinct values. The gain ratio measurement can help reduce the problem by adding $IV(F)$ to penalize attributes with large number of distinct values.

Third, we use a greedy-based strategy to prune the sorted list. Specifically, it iteratively removes the first element in the list and adds it to selected feature list. Then, remove all ancients and descendants of this element in the sorted list. Therefore, the selected features list can be interpreted as a mixture of concepts from different levels of the tree-structure drug ontology. The average time complexity of the prune strategy is $O(n \log n)$, where n is the number of features in the original dataset.

INPUT: Drug train dataset: *trainDataset*, Tree-structure Drug Ontology: *ontologyTree*
OUTPUT: Selected features list: *selectedFeatureList*

GREEDY-BASED-TOP-DOWN-FEATURE-SELECTION

for each *childNode* **in** *ontologyTree*

featureList = TRAVERSE(*childNode*)
orderedFeatureList = SORT(*featureList* , Heuristic_Function)
selected = PRUNE-LIST(*orderedFeatureList* , *ontologyTree*)
selectedFeaturesList.ADD(*selected*)

end

PRUNE-LIST(*orderedFeatureList* , *ontologyTree*)

while HAS_ELEMENT(*orderedFeatureList*) **do**

First_Element = *orderedFeatureList.REMOVE*(first)
selected.ADD(*First_Element*)

for each *element* in *orderedFeatureList*
if ANCIENT(*element*,*First_Element*,*ontologytree*) **or** CHILDREN(element,First_Element,ontologytree) **then**
orderedFeatureList.REMOVE(element)
end if
end
end
return *selected*

Figure 4 Greedy based top-down search strategy for feature selection

C. Machine Learning Model

A Naïve Bayes classifier is used to test the quality of the subset of features we selected. We chose to work with the Naïve Bayes classifier for several reasons. First, with the strong independence assumption, the correlations between

features in the dataset may have a strong impact on the performance of Naïve Bayes classifier. Thus, if we have better performance of the Naïve Bayes classifier, we would probably have selected a better subset of features. Second, with the prior and the likelihood updated dynamically for each training instance, the classifier is robust to errors [15]. This property of Naïve Bayes helps to reduce the impact of the abnormal instances. Third, the outputs have a probability interpretation. Generally speaking, Naïve Bayes classifier often beat other more sophisticated classifiers in practice [16].

D. Evaluation

In our experiments, we compare the performance of our feature selection methods against the performance achieved by feature selection methods without employing drug ontology. Specifically, we choose information gain feature filter and gain ratio separately as the baseline feature filters. The information gain feature filter model ranks the features based on their information gain, and select a number of top ranked features. The gain ratio feature filter uses the same routine, and the difference is that it uses gain ratio as ranking measurement. For the drug ontology based feature selection strategy, we use gain ratio and information gain as heuristic. We use Naïve Bayes classifier as learning model to perform the heart failure readmission prediction. We use the area under the Receiver operating characteristic curve (AUROC) to evaluate performances.

In addition, we also do experiment using a single level of the drug ontology we constructed. Specifically, there are five levels used in the experiment: three legacy VA classes, ingredient level, and RXCUI level. These experiments would help to give us some intuition about the relationship between different drug clustering granularities and readmission prediction.

IV. Experiments and Results

In the experiment, we use prescribed medication during heart failure patients' hospital stay as studying dataset. Specifically, we retrieve all prescribed drug list of heart failure visits from 13 UPMC hospitals between Jan. 1, 2008 and Dec. 31, 2012. The inclusion criteria for heart failure visits are based on primary discharge diagnosis of heart failure as indicated by the following International Classification of Diseases, Ninth Revision—Clinical Modification (ICD-9—CM) Codes of 428 family and 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93. Readmission is defined as re-hospitalization for any cause to any of these 13 UPMC hospitals within 30 days of discharge. There are 15,964 heart failure admissions during the period of study (filtered from 20,933), including 3,831 readmissions (readmission rate: 24%). The features are defined by the local drug codes of the hospital medication system. There are 11,253 distinct local drug codes.

We perform the readmission prediction experiment using Naïve Bayes classifier with nine feature selection methods. Specifically, we use ontology based feature selection using gain ratio as heuristic, which is referred as *ontology_based_selection_GR*; ontology based feature selection using information gain as heuristic, which is referred as *ontology_based_selection_IG*; features only from the first level of legacy VA Class, which is referred as *VA_class_level_1*; features only from the second level of legacy VA Class, which is referred as *VA_class_level_2*; features only from the third level of legacy VA Class, which is referred as *VA_class_level_3*; features only from the ingredient level, which is referred as *Ingredient*; features only from the RXCUI level, which is referred as *RXCUI*; features selected from the local drug codes using gain ratio as measurement, which is referred as *local_drug_code_GR*; features selected from the local drug codes using information gain as measurement, which is referred as *local_drug_code_IG*.

We use General Linear Model (GLM) method [17] employed by SAS® 9.3 (Copyright, SAS Institute Inc.) to analyze feature selection effects on model performance (5-fold cross-validation).

FeatureSelection	AUROC LSMEAN	95% Confidence Limits	
ontology_based_selection_GR	0.5934	0.5839	0.6029
ontology_based_selection_IG	0.5894	0.5799	0.5989
VA_class_level_3	0.5850	0.5755	0.5945
Ingredient	0.5822	0.5727	0.5917
VA_class_level_2	0.5822	0.5727	0.5917
VA_class_level_1	0.5704	0.5609	0.5799
RXCUI	0.5658	0.5563	0.5753
local_drug_code_IG	0.5438	0.5343	0.5533
local_drug_code_GR	0.5000	0.4905	0.5095

Figure 5 Marginal means (least-square means) of AUROCs for models with different feature selection methods

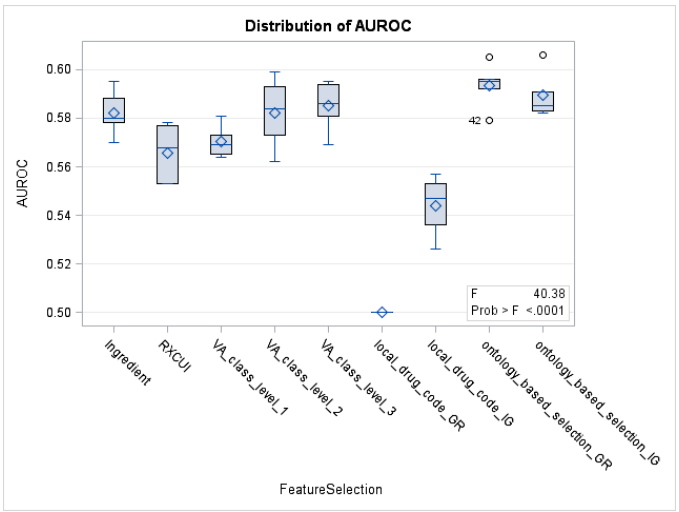

Figure 6 Comparisons of performance of models using different feature selection approaches

The F test shows that the influence of drug ontology based feature selection is extremely significant (F value = 40.38, P<0.0001). Fig. 5 shows the marginal means (least-square

means) [18] of AUROCs for the Naïve Bayes classifier with different feature selection methods. As is shown in Fig. 5, feature subset selected by drug ontology based strategy using gain ratio heuristic outperforms other feature subsets.

Fig. 7 reports pairwise comparisons (Tukey's Studentized Range (HSD) Test) [19] of mean AUROC of models with different feature sets selected with different methods. All of ontological processing methods have significantly improved the performance (p values are all less than 0.05). Moreover, *ontology_based_selection_GR* and ontology_*based_selection_IG* methods are significantly better than *RXCUI* method (differences of AUROC are around 0.02 to 0.03, $p<0.05$). In addition, *ontology_based_selection_GR* is significantly better than *VA_class_level_1* (difference of AUROC is 0.02, $P<0.05$).

Fig. 6 compares the performance of using different feature selection methods. The performance of the drug ontology based feature selection method not only achieves a higher AUROC value but also has a lower variance, which implies the stability of the performance using the method. For the results for each single layer in the drug ontology, the third level of legacy VA Class has a more stable performance. This observation might imply the third level of legacy VA Class can serve a more significant impact on the readmission prediction. With this indication, we can further explore the classification criteria of NDF-RT legacy VA Class level three in order to expand the current tree structure drug ontology, as the current NDF-RT legacy VA Class only covers approximately 93% of the drugs in our original dataset.

Feature Selection Comparison	Difference between means	Simultaneous 95% Confidence Limits		Comparisons significant at the 0.05 level are indicated by ***
ontology_based_selection_GR - local_drug_code_GR	0.0934	0.0717	0.1151	***
ontology_based_selection_IG - local_drug_code_GR	0.0894	0.0677	0.1111	***
VA_class_level_3 - local_drug_code_GR	0.0850	0.0633	0.1067	***
VA_class_level_2 - local_drug_code_GR	0.0822	0.0605	0.1039	***
Ingredient - local_drug_code_GR	0.0822	0.0605	0.1039	***
VA_class_level_1 - local_drug_code_GR	0.0704	0.0487	0.0921	***
RXCUI - local_drug_code_GR	0.0658	0.0441	0.0875	***
ontology_based_selection_GR - local_drug_code_IG	0.0496	0.0279	0.0713	***
ontology_based_selection_IG - local_drug_code_IG	0.0456	0.0239	0.0673	***
local_drug_code_IG - local_drug_code_GR	0.0438	0.0221	0.0655	***
VA_class_level_3 - local_drug_code_IG	0.0412	0.0195	0.0629	***
VA_class_level_2 - local_drug_code_IG	0.0384	0.0167	0.0601	***
Ingredient - local_drug_code_IG	0.0384	0.0167	0.0601	***
ontology_based_selection_GR - RXCUI	0.0276	0.0059	0.0493	***
VA_class_level_1 - local_drug_code_IG	0.0266	0.0049	0.0483	***
ontology_based_selection_IG - RXCUI	0.0236	0.0019	0.0453	***
ontology_based_selection_GR - VA_class_level_1	0.0230	0.0013	0.0447	***
RXCUI - local_drug_code_IG	0.0220	0.0003	0.0437	***

Figure 7 Pairwise comparisons (Tukey) of mean AUROC of models with different feature selection methods

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a method for building a drug ontology based greedy search strategy to do feature dimensional reduction. The method utilizes a tree structure constructed by combining two publicly accessible drug ontology knowledge base systems, namely, RxNorm and NDF-RT. The tree structure drug ontology has six levels. The top three levels are adopted from legacy VA Class level, which can be interpreted as the therapeutic intent of drugs. The fourth level is adopted from RxNorm, which can be viewed as active ingredients of drugs. The fifth level is also adopted from RxNorm, which can be viewed as unified drug identifier. The sixth level is adopted from local hospital medication system, which can be viewed as heterogeneous drug description. For this tree structure, we assume that each node can be interpreted as a feature. A node may have high correlation with its parent and children.

To reduce the influence of high correlation between features, we propose a greedy search strategy to get a list of nodes that have lower correlation with each other, but higher correlation with the class label. We use information gain ratio as measurement for greedy search strategy. The proposed method helps to select a subset of features to predict the probability of heart failure readmission. Naïve Bayes classifier, which is sensitive to high correlated features, is used to perform the prediction task. The experiment results show that our method outperforms feature selection methods without using drug ontology.

As a pilot study of a project who aims to leverage all routinely collected clinical information in EHRs for automatic heart failure readmission risk assessment, this research studies how to utilize drug data only to predict the 30-Day heart failure readmission. In order to achieve a better performance, we would combine other EHR types, such as demographics data, lab data, free-text inpatient reports and so forth, into the train dataset. These heterogamous data are gathered from different sources and the combination task is not trivial. In the future work, we will focus on combining drug data into dataset constructed using data from multiple sources and improve the performance of readmission prediction furthermore.

REFERENCES

[1] "Healthcare Cost and Utilization Project (HCUP) http://hcupnet.ahrq.gov." .

[2] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *ICML*, 2003, vol. 3, pp. 856–863.

[3] J. S. Ross, G. K. Mulvey, B. Stauffer, V. Patlolla, S. M. Bernheim, P. S. Keenan, and H. M. Krumholz, "Statistical models and patient predictors of readmission for heart failure: a systematic review," *Arch. Intern. Med.*, vol. 168, no. 13, p. 1371, 2008.

[4] R. Amarasingham, B. J. Moore, Y. P. Tabak, M. H. Drazner, C. A. Clark, S. Zhang, W. G. Reed, T. S. Swanson, Y. Ma, and E. A. Halm, "An Automated Model to Identify Heart Failure Patients at Risk for 30-Day Readmission or Death Using Electronic Medical Record Data," *Med. Care*, vol. 48, no. 11, pp. 981–988, Nov. 2010.

[5] G. M. Felker, J. D. Leimberger, R. M. Califf, M. S. Cuffe, B. M. Massie, K. Adams Jr, M. Gheorghiade, and C. M. O'Connor, "Risk

stratification after hospitalization for decompensated heart failure.," *J. Card. Fail.*, vol. 10, no. 6, p. 460, 2004.
[6] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J Mach Learn Res*, vol. 3, pp. 1157–1182, Mar. 2003.
[7] M. A. Hall and L. A. Smith, "Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper," in *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference*, 1999, pp. 235–239.
[8] J. Pathak and C. G. Chute, "Further revamping VA's NDF-RT drug terminology for clinical research," *J. Am. Med. Informatics Assoc. JAMIA*, vol. 18, no. 3, pp. 347–348, 2011.
[9] J. Pathak and C. G. Chute, "Analyzing categorical information in two publicly available drug terminologies: RxNorm and NDF-RT," *J. Am. Med. Informatics Assoc. JAMIA*, vol. 17, no. 4, pp. 432–439, 2010.
[10] "RxNorm Overview." [Online]. Available: http://www.nlm.nih.gov/research/umls/rxnorm/overview.html.
[11] "2012AB National Drug File - Reference Terminology Source Information." [Online]. Available: http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/.
[12] O. Bodenreider, F. Mougin, and A. Burgun, *Proceedings of the 13th ISMB'2010 SIG meeting "Bio-ontologies " 2010:140-143. Automatic determination of anticoagulation status with NDF-RT*. .
[13] J. Pathak, S. P. Murphy, B. N. Willaert, H. M. Kremers, B. P. Yawn, W. A. Rocca, and C. G. Chute, "Using RxNorm and NDF-RT to Classify Medication Data Extracted from Electronic Health Records: Experiences from the Rochester Epidemiology Project," *AMIA. Annu. Symp. Proc.*, vol. 2011, pp. 1089–1098, 2011.
[14] J. Mingers, "An empirical comparison of selection measures for decision-tree induction," *Mach. Learn.*, vol. 3, no. 4, pp. 319–342, 1989.
[15] F. Keller, "Naive Bayes Classifiers," *Connect. Stat. Lang. Process. Course Univ. Saarlandes*, 2002.
[16] I. Rish, "An empirical study of the naive Bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 2001, vol. 3, pp. 41–46.
[17] R. C. Littell, *SAS*. Wiley Online Library, 2006.
[18] J. H. Goodnight and W. R. Harvey, *Least squares means in the fixed effects general linear model*. SAS Institute, 1978.
[19] J. W. Tukey, *The problem of multiple comparisons*. Princeton University, 1973.