A Data-Driven Feature Extraction Framework for Predicting the Severity of Condition of Congestive Heart Failure Patients

Costas Sideris[1], Nabil Alshurafa, Mohammad Pourhomayoun, Farhad Shahmohammadi
Lauren Samy, Majid Sarrafzadeh

***Abstract*— In this paper, we propose a novel methodology for utilizing disease diagnostic information to predict severity of condition for Congestive Heart Failure (CHF) patients. Our methodology relies on a novel, clustering-based, feature extraction framework using disease diagnostic information. To reduce the dimensionality we identify disease clusters using co-occurence frequencies. We then utilize these clusters as features to predict patient severity of condition. We build our clustering and feature extraction algorithm using the 2012 National Inpatient Sample (NIS), Healthcare Cost and Utilization Project (HCUP) which contains 7 million discharge records and ICD-9-CM codes. The proposed framework is tested on Ronald Reagan UCLA Medical Center Electronic Health Records (EHR) from 3041 patients. We compare our cluster-based feature set with another that incorporates the Charlson comorbidity score as a feature and demonstrate an accuracy improvement of up to 14% in the predictability of the severity of condition.**

I. INTRODUCTION AND RELATED WORK

Disease diagnostic codes are a valuable resource for classifying and predicting patient outcomes. However, there are more than 15,000 different ICD-9-CM (International Classification of Diseases, Clinical Modification - Version 9) disease and symptom diagnostic codes in use in the US health system and even more in the newer ICD-10 coding scheme. Considering each of the diagnostic codes as a separate feature in a predictive framework can be computationally costly. It is also complicated due to the high variability in disease occurrence frequencies.

Originally, ICD-9-CM and similar coding schemes were introduced to standardize and facilitate hospital billing. Since the introduction and proliferation of Electronic Health Records (EHR), ICD-9-CM have been used in a variety of retrospective clinical studies to identify risk factors for specific outcomes. However, most of this body of work has focused only on specific conditions and prediction of specific outcomes.

Two of the best known comorbidity measures are the Charlson [1] and Elixhauser's Comorbidity [2] Index. The former predicts ten-year mortality for patients based on a range of comorbid conditions such as AIDS, Cancer and Heart Failure. Twenty two conditions are considered in the Charlson Index and are weighted based on their severity with a score of 1, 2, 3, or 6. Elixhauser's measure was developed based on administrative data from the State of California. It takes into consideration a list of 30 comorbidities based on the ICD-9-CM codes. Both measures and their variations have been studied extensively in the context of predicting inpatient death, in-hospital adverse events and readmission risk ([3], [4]).

More recently, several data-driven schemes have been presented that rely on mining administrative data and EHR. Roque et al. [5] describe a framework to discover disease correlations through co-occurrence and map them to biological frameworks. Bauer-Mehren et al [6] created a network from unstructured EHR to examine the efficiency of certain treatments and identify patient cohorts. In a prior effort [7], we described a framework for modeling the severity of condition for CHF patients.

To enable the effective utilization of disease and symptom information in a general classification scheme we propose a flexible, data-driven feature extraction scheme from ICD-9-CM diagnostic codes that can easily adapt to different classification tasks. We examine the efficiency of our scheme in the context of quantifying and predicting CHF patient risk. More specifically, this paper makes the following contributions:

- Designing and developing a flexible, data driven approach for feature extraction from disease diagnostic information.
- Developing a classification scheme for categorizing patients into high and low risk groups

Section II presents the different steps of our methodology and defines six methods for modeling severity of condition for CHF patients. Section III demonstrates the classification accuracy gains of the proposed methodology versus the commonly used Charlson Index and section IV provides a summary of our effort.

II. METHODOLOGY

A. Data

We obtained EHR from the Ronald Reagan UCLA Medical Center between 2005 and 2009. The dataset consists of patients admitted primarily for CHF and related complications. The dataset includes patient demographics (gender, age, race), diagnostic information encoded in ICD-9-CM and hospitalization specific information including blood test results and discharge diagnoses coded as ICD-9-CM codes. The Ronald Reagan dataset contains data from 4406 admissions from 3041 patients over 4 years. This dataset is the motivating factor behind this work as clinicians are increasingly interested in identifying risk factors from EHR data which can provide useful insights into how patients

*This work was not supported by any organization

*All authors are with the Department of Computer Science, University of California, Los Angeles

[1] costas@cs.ucla.edu

978-1-4244-9270-1/15/$31.00 ©2015 IEEE

should be monitored outside the hospital. The dataset contains multivariate signals from 31 sources although not all sources are reported for all patients. These signals have different lengths depending on the length of stay of the patients. All patients however have heart rate and blood pressure signals reported which are measured at an interval of 15 minutes for the duration of their hospital stay. In addition, we used administrative data from the 2012 HCUP NIS sample. The NIS sample contains 7 million discharge records with ICD-9-CM codes.

B. Feature Extraction from ICD-9-CM codes

To reduce the dimensions of diagnostic codes we first identify disease groups with high frequency of co-occurrence. To avoid over-fitting our framework to a specific dataset, we used the HCUP data to compute empirical co-occurrences between ICD-9-CM codes and validated the approach on the Ronald Reagan dataset. These co-occurrence frequencies are modeled through the following Jaccard score:

$$S(A,B) = \frac{P(A \cap B)}{P(A \cup B)} \tag{1}$$

This score will be 0 when two codes never appear together (independent) and 1 when they always appear together. Using the calculated Jaccard score we compute the distance matrix of the codes with the following distance function:

$$D(A,B) = 1 - S(A,B) \tag{2}$$

Subsequently we cluster the ICD-9-CM codes according to their distance using hierarchical clustering based on the minimum variance method [8]. An example of the resulting dendrogram can is provided in Figure 4.

Depending on the height at which the dendrogram is cut, a different number of disease clusters is generated. The clusters are converted into binary features by assigning a 1 if the patient presents a code within that cluster and 0 otherwise. Using an increasing number of such features tends to improve classification accuracy but above a certain number of features over-fitting becomes a concern. To search for the optimal height that maximizes the classification accuracy we propose a greedy methodology. The steps of the methodology are shown in Figure 1. We initialize the search at the maximum height of the dendrogram and we reduce the height (thus increasing the number of features) at every step as long as the classification accuracy on the training dataset increases. At each step we generate a new set of disease features, we select the best subset using correlation-based feature subset selection [9] and we calculate the resulting classification accuracy on the training dataset. We stop the search when the accuracy increases above a certain threshold.

C. CHF Patient Risk Classification

To obtain objective measures of a hospitalized patient's condition we extract six daily threshold-based outcome variables from the collected physiological signals. These variables are used in labeling the severity of condition and have been studied in the context of remote health monitoring

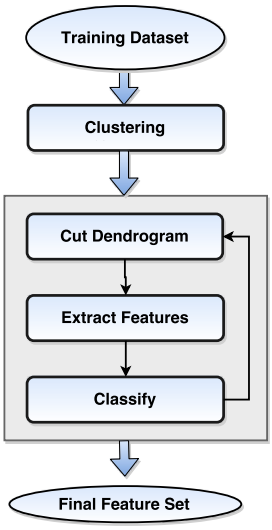


Fig. 1: Steps of the disease feature extraction methodology

of CHF patients [10]. The outcome variables have a value of 1 when a physiological measurement is out of the predefined acceptable range and 0 otherwise. The outcomes are also ranked as medium or high priority depending on the degree of deviation from the defined thresholds. The complete list of outcome variables we extract from heart rate, blood pressure signals can be seen in table I:

TABLE I: Daily Indicators

Outcome Variables	Description	Priority
I_1	Heart Rate > 120 bpm	High
I_2	Heart Rate < 50 bpm	High
I_3	100 < Heart Rate < 119 bpm	Medium
I_4	Systolic BP < 80 mmHg	Medium
I_5	Systolic BP > 160 mmHg	Medium
I_w	$3*I_1+3*I_2+I_3+I_4+I_5$	

As provided in Table I, we also include a weighted sum of the outcome variables (I_w). The weighted sum is calculated based on their priority [10]. We extract such outcome variables from the Ronald Reagan dataset from every single admission record on a daily basis and calculate the average daily value per type. We exclude from these calculations the first day of admission. Physiological signals during the first day of hospitalization are used as well to help predict patient severity of condition. Statistical features are extracted from the first day vitals of each patient to be used as predictive features. The following features are computed during the first day of hospitalization from heart rate, systolic & diastolic blood pressure and weight: mean, max, min, range, standard deviation during the first day of hospitalization. After processing the dataset, a total of 3057 valid records were identified from 1948 patients.

Furthermore, we calculate a binary outcome variable by thresholding. For each of the examined alerts we try to predict whether a patient's vitals will generate an average of 0 or more number of alerts each day. The transformed prediction

problems now distinguish between low risk individuals (i.e no daily deviation from thresholds) and high risk, for each of the outcomes.

For each of the binary outcome variables, I_1, I_2, ..., I_w, we extract the most correlated physiological features. Subsequently, we compare the classification accuracy using those features together and the disease features generated by our methodology against the same physiological features and the Charlson comorbidity Index. The Charlson index is calculated in the basis of the Quan revision of Deyo's ICD-9-CM mapping [11]. For each of the two compared methods and for each of the subproblems we train support vector machine classifiers [12] and perform 10-fold cross-validation evaluation.

III. RESULTS

A. Feature Extraction

To select the best number of disease features, we search for the number that maximizes the classification accuracy on the training dataset. Figure 2 displays the classification accuracy on the training dataset for a specific cross-validation fold. The dashed line represents the chosen number of features as a result of the search process (on the training dataset). We also present the resulting accuracy on the testing dataset. It is easy to see that our greedy algorithm locates a good local maximum that tends to maximize the classification accuracy on the test dataset.

Table II summarizes the average and maximum difference between our framework and using Charlson's index. It also displays the number of disease features that result in the maximum classification accuracy improvement.

TABLE II: Optimal Number of Disease Features

Alert	Mean Diff.	Max Diff.	No. Features/Clusters (Max Diff.)
I_1	4.54%	10.53%	21
I_2	6.94%	14.29%	21
I_3	2.42%	4.23%	22
I_4	1.51%	10.96%	23
I_5	0.24%	5.49%	10
I_w	4.71%	8.82%	21

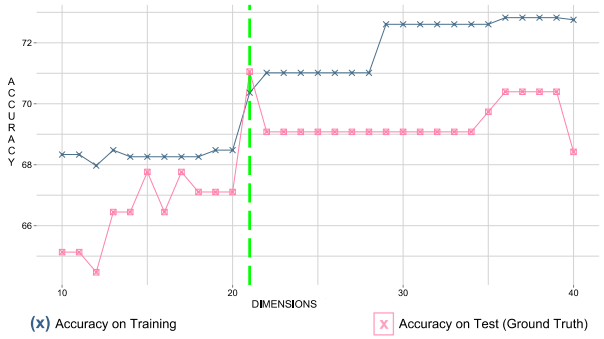
B. Severity of Condition Classification

The results from the two methods for predicting severity of condition are shown in Table III.

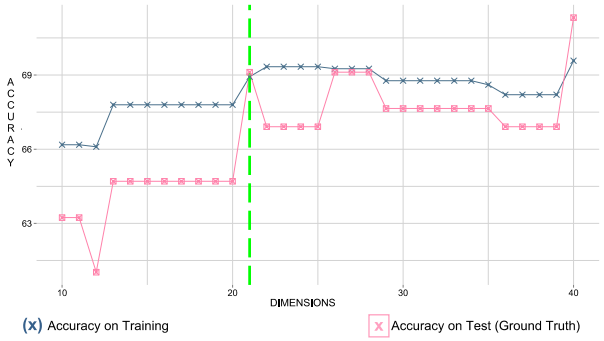
TABLE III: Accuracy Comparison

Alert	Accuracy		Our Framework		Charlson Index	
	Ours	Charlson	TPR	TNR	TPR	TNR
I_1	70.72	66.18	64.21	77.24	59.74	72.63
I_2	58.57	51.63	52.65	64.49	53.06	50.20
I_3	73.15	70.73	67.31	79.00	64.31	77.15
I_4	65.48	63.97	71.78	59.18	72.74	55.21
I_5	69.39	69.15	63.66	75.12	61.10	77.20
I_w	67.87	63.16	54.71	81.03	52.94	73.38

Our proposed methodology significantly improved classification accuracy as well as True Positive and True Negative Rates for each of the subproblems. An improvement is achieved in every task examined and ranges from 0.24% to 7% on average and up to 14.3% in certain folds. Figure 3 summarizes these findings.



(a) I_1



(b) I_w

Fig. 2: Classification Accuracy vs # disease clusters for the I_1 alert type (Heart Rate > 120 bpm) (a) and I_w alert type (Weighted Alert) (b). The dashed green line highlights the selected number of disease features by our algorithm.

C. Cluster Analysis

Detailed results of the features/ clusters selected for each outcome are beyond of the scope of this effort. However, to provide the reader with insight into the generated disease features we produce clusters by manually choosing a height of 1.5 to cut the dendrogram (Figure 4). This results in 13 disease features where one of them (D1) is significantly larger than the rest and can be explained as a grouping of diseases with no significant co-occurence frequencies. We populate the description of each group by text mining the ICD-9-CM descriptions to obtain the most frequently mentioned phrases. Table IV describes the most representative diagnoses of each of the generated clusters for the aforementioned height of dendrogram.

IV. CONCLUSIONS

We have presented a novel data-driven framework to extract predictive features from disease and symptom diagnostic codes. Our greedy optimization methodology automatically locates a suitable number of cluster-based features that significantly improves outcome prediction accuracy. We have successfully applied this framework to predict a CHF patient's severity of condition and showed significant gains

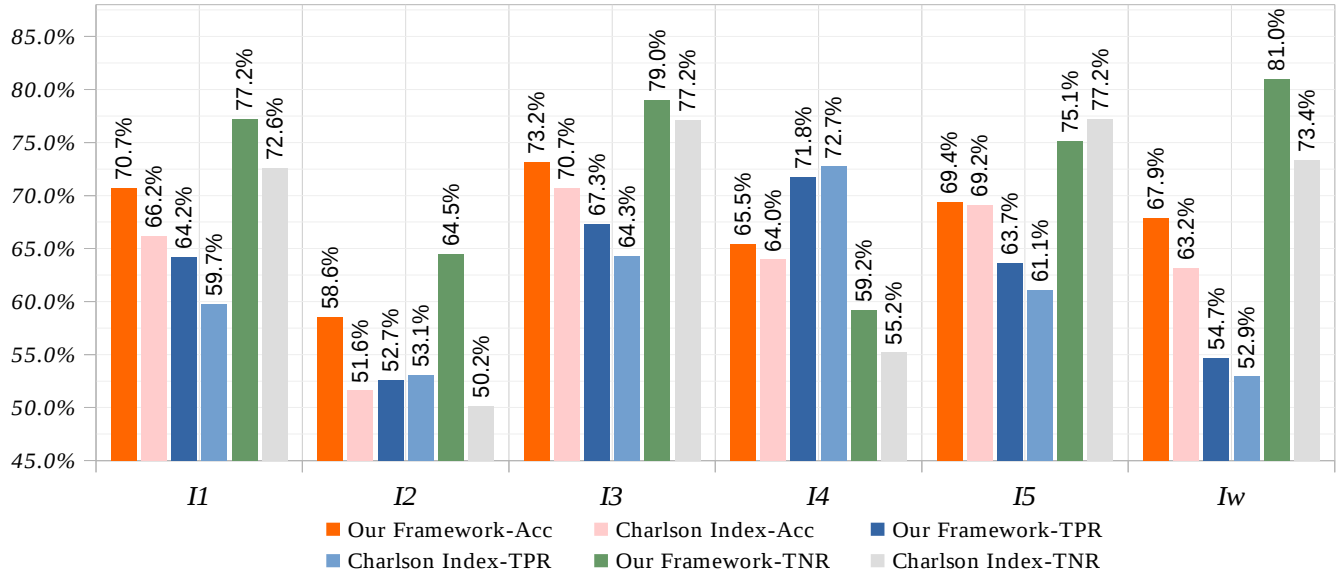


Fig. 3: Improvements in the classification accuracy.

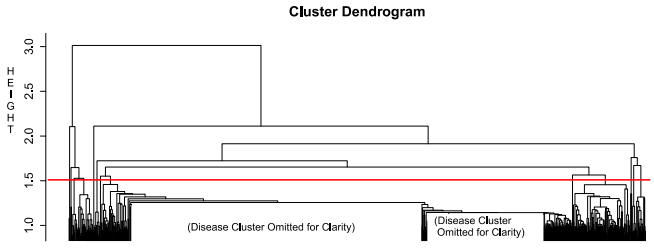


Fig. 4: Disease Dendrogram with the selected clusters.

compared with the commonly used Charlson Index. This approach can be extended to other chronic diseases by targeting different risk indicators as outcome variables.

TABLE IV: Disease Clusters for Dendrogram Height = 1.5

Code	Description
D_1	Remaining Codes*, mention of malignant neoplasms
D_2	Heart failure, Valve disorders, Acute kidney failure, Respiratory failure
D_3	Congestive Heart Failure, Coronary Disease, Tobacco Use
D_4	Cerebrovascular disease, Malignant neoplasm, Substance Abuse
D_5	Diabetes II or unspecified, Aherosclerosis, Peripheral circulatory disorders
D_6	Diabetes II, nephropathy, retinopathy, neurological complications
D_7	Diabetes II, Liver disease, Cirrhosis, Digestive Complications
D_8	Chronic kidney disease , Renal Disease (End Stage)
D_9	Neoplasms of Hematopoietic cells , Leukemia
D_{10}	Diabetes I (juvenile type), related complications
D_{11}	Pressure ulcer
D_{12}	Sepsis, Septic Shock
D_{13}	Dementia,Cerebral Atherosclerosis

REFERENCES

[1] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation," *Journal of chronic diseases*, vol. 40, no. 5, pp. 373–383, 1987.

[2] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey, "Comorbidity measures for use with administrative data," *Medical care*, vol. 36, no. 1, pp. 8–27, 1998.

[3] D. A. Southern, H. Quan, and W. A. Ghali, "Comparison of the elixhauser and charlson/deyo methods of comorbidity measurement in administrative data," *Medical care*, vol. 42, no. 4, pp. 355–360, 2004.

[4] J. F. Farley, C. R. Harley, and J. W. Devine, "A comparison of comorbidity measurements to predict healthcare expenditures." *The American journal of managed care*, vol. 12, no. 2, pp. 110–119, 2006.

[5] F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Søeby, S. Bredkjær, A. Juul, T. Werge *et al.*, "Using electronic patient records to discover disease correlations and stratify patient cohorts," *PLoS computational biology*, vol. 7, no. 8, p. e1002141, 2011.

[6] A. Bauer-Mehren, P. LePendu, S. V. Iyer, R. Harpaz, N. J. Leeper, and N. H. Shah, "Network analysis of unstructured ehr data for clinical research," *AMIA Summits on Translational Science Proceedings*, vol. 2013, p. 14, 2013.

[7] C. Sideris, N. Alshurafa, B. Shahbazi, M. Sarrafzadeh, and M. Pourhomayoun, "Using electronic health records to predict severity of condition for congestive heart failure patients," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 2014, pp. 1187–1192.

[8] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[9] M. A. Hall and L. A. Smith, "Practical feature subset selection for machine learning," 1998.

[10] M. Pourhomayoun *et al.*, "Multiple model analytics for adverse event prediction in remote health monitoring systems," *Healthcare Innovation Conference (HIC), 2014 IEEE*, pp. 106–110, 2014.

[11] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.-C. Luthi, L. D. Saunders, C. A. Beck, T. E. Feasby, and W. A. Ghali, "Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data," *Medical care*, pp. 1130–1139, 2005.

[12] J. Platt *et al.*, "Sequential minimal optimization: A fast algorithm for training support vector machines," 1998.