Sequence Based Prediction of Hospital Readmissions

Surabhi Agrawal
Center for Data Science
Institute of Technology
University of Washington
Tacoma
agraws@uw.edu

Chun Pan Hon
Center for Data Science
Institute of Technology
University of Washington
Tacoma
darrencp@uw.edu

Swati Garg
Center for Data Science
Institute of Technology
University of Washington
Tacoma
swatig1@uw.edu

Aadarsh Sampath
Center for Data Science
Institute of Technology
University of Washington
Tacoma
aadsam@uw.edu

Shanu Sushmita
Center for Data Science
Institute of Technology
University of Washington
Tacoma
sshanu@uw.edu

Martine De Cock*
Center for Data Science
Institute of Technology
University of Washington
Tacoma
mdecock@uw.edu

ABSTRACT

In this contribution we explore the use of an N-gram based approach to make predictions about the next event in a sequence of hospital admissions. To predict the value of a variable in the next admission (e.g. *length of stay*), we use only the sequence of values of the same variable for all hospital admissions of the same patient so far (e.g. *the length of all previous hospital stays*) and no other information from the discharge record. We validate our method on inpatient data from the California Office of Statewide Health Planning and Development (OSHPD), for the prediction of three kinds of variables: whether the patient will be readmitted within 30 days, and what the associated length of stay and cost of the next hospital admission will be. Our preliminary results show that in all cases simple 4-gram, 5-gram and 6-gram methods make more accurate predictions than the majority baseline method, by a margin that depends on the problem at hand as well as on the cohort, i.e. the patient population as a whole ("ALL-Cause") versus a congestive heart failure (CHF) cohort.

1. INTRODUCTION AND BACKGROUND

Sequences are an important type of data which occur frequently in several scientific, medical, security, and other domains [8]. Sequence data mining provides helpful tools for exploring useful knowledge hidden in large sequence datasets. Many prediction problems in clinical care settings are inherently sequence prediction problems: given a sequence of previous patient encounters (e.g. hospital visits) the task is to predict one or more aspects about future encounters with the same patient.

The problem of hospital readmissions is very severe in the U.S. and currently 1 in 5 patients is readmitted to the hospital within 30 days of being discharged. Measures that can reduce the number of unnecessary, lengthy and costly hospital readmissions are therefore very valuable [2]. The ability to predict them accurately provides many benefits for accountable care, now a global issue and foundation for the U.S. government mandate under the Affordable Care Act. While predicting 30-day readmission has been identified as one of the key problems for the healthcare domain, not many solutions are known to be effective [4]. Furthermore, uncertainty in length of hospital stay is a major deterrent to effective scheduling for admission of elective patients. A model to predict the Length of Stay (LOS) for hospitalized patients can be an effective tool for healthcare providers, as it will enable early interventions to prevent complications, among other things [1]. However, the ability to risk stratify for LOS based on patient admission and hospital characteristics is limited, and particularly challenging for congestive heart failure (CHF) patients. Readmissions and prolonged hospital stay act as substantial contributors to rising healthcare costs [3]. Alongside of predicting *30-day readmission* and *LOS*, in this contribution we also investigate algorithm performance for forecasting the *cost* of hospital admissions. Previously proposed cost prediction models were primarily focused on 'general' healthcare costs as apposed to hospital admissions, and were often rule based and regression models. Additionally, the development of healthcare cost prediction models using data mining techniques has been more recent (e.g. [7]). It is a common practice to make predictions about the *next* hospital admission using the discharge record of the *current* admission. Leveraging the temporal dimension of the entire sequence of prior hospital admissions has remained largely unexplored so far.

In this contribution we explore the use of a statistical N-gram based approach to predict variables about the next event in a sequence of admissions of the same patient. These models exploit probability distributions over sequences of events and have been successfully utilized in several domains (e.g., information retrieval [6]). We use an N-gram approach to predict whether the patient will be readmitted within 30 days, and what the associated length of stay and cost of the next hospital admission will be. The only input features we use are the values of these same three variables (30-Day, LOS and cost) for all prior admissions of the same patient. We validate our method on inpatient data from the California Office of Statewide Health Planning and Development (OSHPD), on the patient population as a whole (ALL-Cause), as well as on a congestive heart failure (CHF) cohort. To the best of our knowledge, our study is the first to explore the potential of next event prediction in symbolic sequences extracted from a large inpatient dataset.

2. METHODS AND RESULTS

We requested non-public OSHPD data for the years 2009-2013. The dataset is a collection of records in tabular format with each row corresponding to one hospital discharge record of one patient. After applying a series of data preprocessing steps (see [5] for a

*Guest professor at Ghent University

BCB '16 October 02-05, 2016, Seattle, WA, USA

ACM ISBN 978-1-4503-4225-4/16/10.

DOI: http://dx.doi.org/10.1145/2975167.2985647[dx.doi.org]

Problem		# Seq	Sequence Length			Symbols
			Min	Max	Avg (SD)	
ALL-30	train	2,538,134	1	333	2.5 (3.2)	0,1
ALL-30	test	281,255	1	174	2.5 (3.2)	0,1
CHF-30	train	530,594	1	333	3.9 (4.6)	0,1
CHF-30	test	58,814	1	174	3.9 (4.6)	0,1
ALL-LOS	train	2,538,134	1	333	2.5 (3.2)	1,2,3,4,5,6
ALL-LOS	test	281,255	1	174	2.5 (3.2)	1,2,3,4,5,6
CHF-LOS	train	448,067	1	255	3.9 (4.4)	1,2,3,4,5,6
CHF-LOS	test	58,814	1	174	3.9 (4.6)	1,2,3,4,5,6
ALL-COST	train	2,125,302	1	255	2.5 (3.2)	1,2,3,4
ALL-COST	test	235,879	1	155	2.5 (3.1)	1,2,3,4
CHF-COST	train	448,067	1	255	3.9 (4.4)	1,2,3,4
CHF-COST	test	49,630	1	164	3.9 (4.5)	1,2,3,4

Table 1: Statistics of the train and test sequence datasets for both cohorts ("ALL-Cause" and "CHF") for the 3 prediction problems under study, i.e. whether the next admission will be within 30 days or not, and what the length and cost of the next admission will be.

description of similar steps) we extracted sequences for our predictions problems as shown in Table 1. For instance, for the 30-day readmission problem in the general population (ALL-30), our preprocessed dataset contains 2,819,389 sequences. Each sequence corresponds to a unique patient. The symbols in the sequence are 1s and 0s, denoting respectively that the next admission of the patient was within 30 days of the previous admission, or not. For instance, the sequence "1,0,0,1,0,1,1" means that the patient had 7 hospital readmissions, 4 of which were within 30 days. Out of the 2,819,389 sequences we set aside approx. 10% for testing (281,255 sequences) and used the rest (2,538,134 sequences) for training, i.e. to construct the N-gram dictionary (see below).

The other sequence datasets in Table 1 were constructed in a similar way. The sequences in the CHF datasets are limited to those of patients who have CHF as a primary or secondary diagnosis in at least one of their records[1]. For LOS we created 6 symbols, corresponding to the discretization used in the well known LACE index [9], i.e. 1, 2, 3, [4-6], [7-13], and 14+ days. For cost we created 4 symbols, each corresponding to a quartile in the data[2]. We add the symbol -1 in front of all the training and test sequences to mark the beginning of a sequence. To predict the next value for a given test sequence using N-grams, probabilities are estimated during the training phase, which are then used during the test phase to predict the next most likely symbol.

Training Phase: The frequencies of all possible grams (1-gram, 2-gram, etc.) in the training sequences are computed and stored in a dictionary. E.g. the training sequence "-1, 2, 1, 0, 3, 2, 0" will lead to the following frequency increments: (-1):1, (2):2, (1):1, (0):2, (3):1, (-1,2):1, (2,1):1, (1,0):1, (0,3):1, (3,2):1, (2,0):1, (-1,2,1): 1, (2,1,0):1, (1,0,3):1, (0,3,2):1, (3,2,0):1, etc.

Testing Phase: Let L denote the length of the test sequence. In the N-gram approach, we take the last M symbols of the test sequence, with M = min(L,N-1), and we complete this to the most probable M+1 gram. I.e., given the last M symbols $(a_1, \ldots, a_M)$, we look up the most frequent M+1 gram $(a_1, \ldots, a_M, b)$ and predict b as the next symbol in the sequence. E.g., the 5-gram approach will consult all 5-grams from the dictionary whose prefix matches the last 4 symbols of the test sequence. It will only resort to grams of a smaller length if the test sequence has less than 4 symbols, and it will never consult any grams of length > 5.

The results for all prediction tasks are presented in terms of accuracy in Table 2. We tested the accuracy of N-gram models from N = 1 to 6, and contrasted their performance with random guessing as well with a majority baseline technique that systematically predicts the most frequently occurring symbol in the training data. For instance, for 30-day readmission, the majority method always predicts 0 as the next symbol, since the majority of readmissions is not within 30 days. Three key observations can be made: (1) the N-gram models consistently perform better than the majority baseline technique for all prediction tasks, but marginally better for CHF-30-Day and CHF-LOS prediction; (2) accurately predicting variables related to the next hospital event using an N-gram approach appears easier for the patient population as a whole (ALL) than the specific cohort (CHF); (3) increasing longitudinal information (longer sequence, hence bigger N) not necessarily seems to further improve the performance. The performance of the 4-gram models is comparable to, and sometimes even better than, that of the 5- and 6-gram models.

[1]Using the ICD-9-CM codes for CHF: 398.91 and 428.XX.

[2]Cost ranges (in dollars) for ALL-Cause: 0-19,080, 19,080-35,690, 35,690-70,450, 70,450+, and for CHF: 0-27,660, 27,660-49,810, 49,810-98,500, 98,500+.

Method	ALL 30	CHF 30	ALL LOS	CHF LOS	ALL COST	CHF COST
Random	50.00	50.00	16.67	16.67	25.00	25.00
Majority	76.93	69.07	22.22	25.47	25.79	26.08
2-gram	76.93	69.07	25.63	25.47	33.26	29.87
3-gram	77.01	69.52	26.20	25.93	33.79	30.05
4-gram	77.02	69.66	26.22	**26.03**	33.92	**30.18**
5-gram	77.10	**69.74**	**26.25**	**26.03**	33.95	30.12
6-gram	**77.17**	**69.74**	26.19	25.78	**33.97**	30.16

Table 2: Performance comparison for 30-Day, LOS and cost of next hospital admission prediction task. The best results are highlighted.

3. CONCLUSION

Being able to stratify patients according to 30-day readmission risk, anticipated length and cost of stay can guide clinicians in discharge planning and intervention recommendation, leading to an increase of quality of care, and a decrease of healthcare cost. In this contribution we explored the use of a statistical N-gram based approach to predict variables about the next event in a sequence of admissions of the same patient. Our preliminary results show that simple N-gram methods make more accurate predictions than the majority baseline method, indicating that there is value in taking sequential information into account. In future work we plan to enrich the simple symbolic sequences considered in this study with features from hospital discharge records.

4. REFERENCES

[1] P. R. Hachesu, M. Ahmadi, S. Alizadeh, and F. Sadoughi. Use of data mining techniques to determine and predict length of stay of cardiac patients. *Healthc Inform Res*, 19(2):121–129, Jun 2013.

[2] A. Hines, M. Barrett, H. Jiang, and C. Steiner. Conditions with the largest number of adult hospital readmissions by payer. *HCUP Statistical Brief*, 172, 2011.

[3] S. F. Jencks, M. V. Williams, and E. A. Coleman. Rehospitalizations among patients in the Medicare fee-for-service program. *New England Journal of Medicine*, 360(14):1418–1428, 2009.

[4] K. Ottenbacher, P. Smith, S. Illig, R. Linn, R. Fiedler, and C. Granger. Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke. *Journal of Clinical Epidemiology*, 54(11):1159–1165, 2001.

[5] M. Pereira, V. Singh, C. P. Hon, T. G. McKelvey, S. Sushmita, and M. De Cock. Predicting future frequent users of emergency departments in California state. In *Proceedings of MAHA*, 2016.

[6] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings ACM SIGIR*, pages 275–281, 1998.

[7] S. Sushmita, G. Khulbe, A. Hasan, S. Newman, P. Ravindra, S. B. Roy, M. De Cock, and A. Teredesai. Predicting 30-day risk and cost of "all-cause" hospital readmissions. In *Proceedings of HIAI*, pages 453–461, 2016.

[8] J. Yang, J. McAuley, J. Leskovec, P. LePendu, and N. Shah. Finding progression stages in time-evolving event sequences. In *Proceedings of WWW*, pages 783–794, 2014.

[9] B. Zheng, J. Zhang, S. W. Yoon, S. S. Lam, M. Khasawneh, and S. Poranki. Predictive modeling of hospital readmissions using metaheuristics and data mining. *Expert Systems with Applications*, 42(20):7110–7120, 2015.