Journal of Big Data

RESEARCH Open Access

HPCC based framework for COPD readmission risk analysis

Piyush Jain[1*], Ankur Agarwal[1], Ravi Behara[2] and Christopher Baechle[3]

*Correspondence:
pjain2012@fau.edu
[1] Department of Computer & Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL, USA
Full list of author information is available at the end of the article

Abstract

Prevention of hospital readmissions has the potential of providing better quality of care to the patients and deliver significant cost savings. A review of existing readmission analysis frameworks based on data type, data size, disease conditions, algorithms and other features shows that existing frameworks do not address the issue of using large amounts of data that is fundamental to readmission prediction analysis. Available patient data for readmission risk analysis has high dimensionality and number of instances. Further, there is more new data produced everyday which can be used on a continuous basis to improve the prediction power of risk models. This study proposes a High Performance Computing Cluster based Big Data readmission risk analysis framework which uses Nave Bayes classification algorithm. The study shows that the over-all evaluation time using Big Data and a parallel computing platform can be significantly decreased, while maintaining model performance.

Keywords: COPD readmission, Prediction, Nave Bayes, HPCC, Big Data

Introduction

The Affordable Care Act enacted in 2010 has resulted in an increased focus on reducing hospital readmissions there by improving quality of care and providing cost savings opportunities [1]. In 2013, hospital readmissions cost more than $26 billion annually [2]. Recently, according to the Institute for Health-care Improvement, 76% of the readmissions which were preventable costed $17 billion [1, 2].

In response to the hospital readmission problem, the Affordable Care Act (ACA) established the hospital readmission reduction program (HRRP) that requires the Center of Medicare and Medicaid Services (CMS) to penalize hospitals which are over a certain threshold of readmissions as compared to others. This penalty is applied by reducing payments to IPPS (Inpatient Prospective Payment System) for three DRG (Diagnosis Related Groups) of cardiac and pulmonary conditions with the highest avoidable readmission [3]. This study focuses on chronic obstructive pulmonary disease (COPD), which is one the leading causes of disability and mortality worldwide [4]. COPD has a significant readmission problem, with 20% of Medicare beneficiaries hospitalized for COPD are readmitted within a 30-day period [5, 6].

Hospital readmissions are consequence of commissions and omissions made during the diagnosis and treatment of patients during the initial hospitalization, or a result of a poorly managed care transition between hospital and other care facilities or the community [7–10]. As a result, targeted intervention programs can be established for high risk patients to reduce the risk of readmission by providing out-of-hospital care thereby improving quality of care and providing better care coordination.

The widespread use of EHR systems has produced vast amounts of data such as discharge summaries in the form of clinical text which can be used to predict high risk patients. Several attempts have been made to create such prediction models [1, 11–17], but all of the models show one of following shortcomings: (1) the prediction model is not created incorporating practical needs of using Big Data [1], (2) the model uses insurance claims data which is not always accurate and would not be available in real-time clinical settings [11, 12], (3) the model is generalized for all-cause readmission and does not take into consideration intricacies specific to each disease condition [1], and (4) the prediction models are not created for continuous learning base on the availability of new data.

The model in this study addresses all the shortcomings and creates a Naive Bayes based classification model for COPD related readmissions. The Naive Bayes implementation allows the system to be highly scalable and non-sensitive to irrelevant data. Also, newly acquired patient data can be incrementally utilized to train the model which will allow the model to be relevant over time. The model is designed on a high performance computing cluster (HPCC) cluster using enterprise control language (ECL) which allows the model to be used in real life scenarios where large number of new discharge summaries can be continually utilized to create a better model.

Literature review

Diagnosis related

Healthcare related information has seen a tremendous growth in recent times [18]. The large amount of data available is being used in high performance computing architectures in various applications such as extracting high-cost patients, analyzing and predicting hospital readmissions [19–21], triage where the risk of complications is estimated [22], predicting decompensating risk of a patient, predicting adverse events well before they occur [23], and predicting diseases affecting multiple organ systems [24]. Healthcare data fulfills all 4 definitions of Big Data: volume, variety, veracity, and velocity [25]. The volume of healthcare data has continued to grow in recent times and every day more and more patients, healthcare institutions, and health insurance companies are adopting electronic operations and produce data in variety of areas from gene data to discharge summaries. In order to utilize the data efficiently, the input and output to the systems should be fast and less time consuming. Big Data applications include medical R&D efficiency [26], Medicare fraud detection [27], reducing hospital readmission [28], and wellness predictive modeling [29]. This paper studies readmission risk prediction for COPD patients from a Big Data perspective.

After the HRRP was established, the readmission problem has been an important focus of the healthcare industry. It has affected all healthcare agencies including CMS in achieving a better overall healthcare economy, hospitals through a reduction of the inpatient prospective payment system (IPPS), and most importantly for patients and family members as this reduces the chances of getting readmitted thereby improving outcomes. Due to these impacts, significant research has been conducted in the area of reducing readmission from various aspects such as demographics and disease types. This section discusses some of the noteworthy studies conducted in the area of reducing readmissions. This study did not limit its focus to a certain type of disease, thereby not limiting the potential advantages and drawbacks which can be further understood in the context of other disease. The studies in this literature review were analyzed and compared based on the data type, data size, disease conditions, algorithms and other features which could be valuable for addressing the readmission problem.

Mehdi et al. [1] conducted a study on all-cause risk of 30-day readmission on 323,813 inpatient stays and used Neural Network model with 1667 features in various feature categories including encounter reason and hospital problems which yielded a precision or positive predictive value (PPV) of 0.24 which was 20% higher than LACE (length of stay, acuity of admission, comorbidities and emergency department visits) which is industry standard. The study also performed a basic cost analysis showing savings as a function of intervention success rate. The study performed by Mohsen et al. [15] on readmission problem for CHF (congestive heart failure) showed a reduction of 18.2% in re-hospitalizations with a cost savings of 3.8%. The study was performed using Logistic Regression with 3888 binary variables extracted from the patient visit data of 1172 hospital visits.

The study performed by Futoma et al. [13] was based on the dataset gathered from the New Zealand Ministry of Health with 3.3 million hospital readmissions between 2006 and 2012. The study showed that the analysis can be used for US healthcare data as well. This study was performed using logistic regression, random forests and support vector machine for COPD, CHF, Pneumonia, THA/TKA and AMI. The data size for COPD was 31,457 which showed an AUC of 0.711. The study performed by Issac et al. [11] was applicable to HF, Acute Myocardial Infarction, Pneumonia or COPD. The data of 7200 records was gathered from administrative records of Veteran Health Administration which correspond to 2985 distinct adult patients. The study shows that PHSF (Phase-Type Random Forests) works better than Random Forests, SVM, Logistic Regression or Neural Network.

Danning et al. [12] use claim based data to predict 3-day readmission using standardized billing code for Chronic Pancreatitis. The study utilized data of 26,091 admissions from John Hopkins Hospital and 16,194 admissions from Bayview Medical Center showing AUC of more than 0.65. In an another study by Amarasingham et al. [14], the use of data directly from multi-condition EHR system across 7 Hospital systems was performed for a patient record set of 39,604. The model from the study was compared with acute decompensated heart failure registry (ADHERE) model and CMS models, and was shown to perform better. The study also derived that claims based models are not efficient, because as claims data are gathered at a very late stage and the data might not be as useful by that time.

The cost sensitive study performed by Christopher et al. [30] consisted of dataset of 1248 patient discharge summaries and a total of 5429 features were extracted based using bag-of-words. The dataset was somewhat imbalanced with class distribution of 14.32% as positive class (readmission) and 85.68% as negative class (non-readmission). The classification algorithms chosen for this study were Nave Bayes (NB), Random Forest (RF), Support Vector Machines (SVM), k-Nearest Neighbors (kNN), C4.5, Bagging with REPTree, and Boosting with Decision Stump. The study shows that by including cost factor in classification, the CMS penalties can be reduced.

In an another study performed by Christopher et al. [31], hospital readmission dataset is shown to be imbalanced in nature, and using that data to create models does not provide efficient solutions. So they proposed a method which uses an ensemble of topic learners to leverage data from multiple hospitals and sources. This study was performed on a dataset of 62,714 instances from 16 hospitals with a total of 7112 extracted features in the corpus.

The study used Nave Bayes (NB), k-nearest neighbors (kNN), linear regression (LR), and support vector machine (SVM) classification algorithms. The results showed that hospitals that implement latent topic ensemble learners using Nave Bayes reduce readmissions and CMS penalties when compared those using other known methods.

Although some of the works reviewed in this section are better at predicting readmission than others, overall they lack two very important aspect of data analysis in real world setting: frameworks suitable for increasing amounts of data, and the to handle new data that is being extracted on a daily basis. Big Data will play a very important role if the hospital readmission prediction system is to be used in a real world setting where all current patients are being marked for readmission probability as they are treated and the feature sets are also being updated as new data is extracted. This study is more focused towards the aspect of implementation using Big Data and using models in such a context.

HPCC related

The Big Data platform selected for this study is the high performance computing cluster (HPCC) systems. It is also known as data analytics supercomputer (DAS). The HPCC systems is an open-source Big Data software architecture developed by LexisNexis Risk Solutions. It provides the architecture which is implemented on commodity computing clusters to deliver high speed output using Big Data [32, 33]. HPCC systems use commodity hardware as processing clusters using high-speed network which ensure that the real time data analytics of readmissions in healthcare is cost-effective. The HPCC system architecture provides high redundancy and availability as the systems store file part replicas on several nodes which makes sure that, in the event of a failure, the data can be provided with no issues.

HPCC systems architecture provides some pre-built tools to create and manage a Big Data platform with ease and efficiency [34]. The tools include administrative tools which allows easy configuration is a cluster environment and job monitoring to keep track of all job units being processed. It also provides some extension modules for natural language parsing, machine learning, and data encryption which can be easily used in the healthcare domain for predicting readmissions using patient discharge summaries [35]. HPCC systems also provides an easy to use Big Data architecture driven declarative language known as enterprise control language (ECL). The ECL compiler is cluster-aware which automatically optimizes the code for parallel processing.

HPCC systems provide many advantages when compared with its alternative Hadoop which is based on Googles Map Reduce paradigm [36]. HPCC systems uses three types of parallelism: data parallelism, pipeline parallelism and system parallelism whereas

Hadoop only uses one type of parallelism [36, 37]. According to a study by Seref et al. [25], for the same 400-node system hardware configuration, HPCC took 6 min and 27 s whereas Hadoop took 25 min and 28 s which shows that HPCC systems is designed very efficiently and provides optimum performance for the same hardware [25]. HPCC systems use ROXIE which was built on architecture of random access, low latency and high concurrency which provides real time query output, but Hadoop does not provide real time processing. One of the distinguishing features of HPCC is its suite monitoring services and tools to ensure high availability. This suggests that HPCC systems can enable use of Big Data in healthcare more effectively and efficiently.

HPCC systems is being used in a wide range of applications including parameter estimation for improving machine learning models [38] and cyber security analytics [39–41]. The healthcare applications utilizing HPCC platforms show great potential of HPCC in this domain as well, as it covers a wide range of applications detecting organized crime in healthcare using social network analytics [42].

Methodology

Ethics

This study was conducted using de-identified patient health records including clinical notes and discharge summaries taken from multiple hospital EHR systems in South Florida. This study was approved by the Institutional Review Board.

Data preparation

Discharge summaries and clinical notes that included demographics, history, treatment plan, discharge comments, equipment used, suggested therapies and co-morbidity information were extracted for 191,205 hospital admissions from various hospitals across South Florida. The data gathered was from 2009 to 2015. Figure 1 also shows admission record count by year. The data is gathered from a variety of care centers including but not limited to Skilled Nursing Facility, Rehabilitation centers, Independent Hospitals, Hospital Associations, Nursing Homes and Long-Term care centers. This ensures that this study captures readmission from various types of care facilities. Figure 2 shows the total number of records from all hospitals. The care

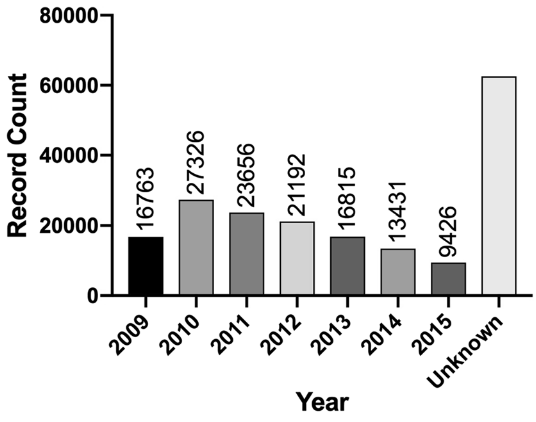

Fig. 1 Discharge summary count by year

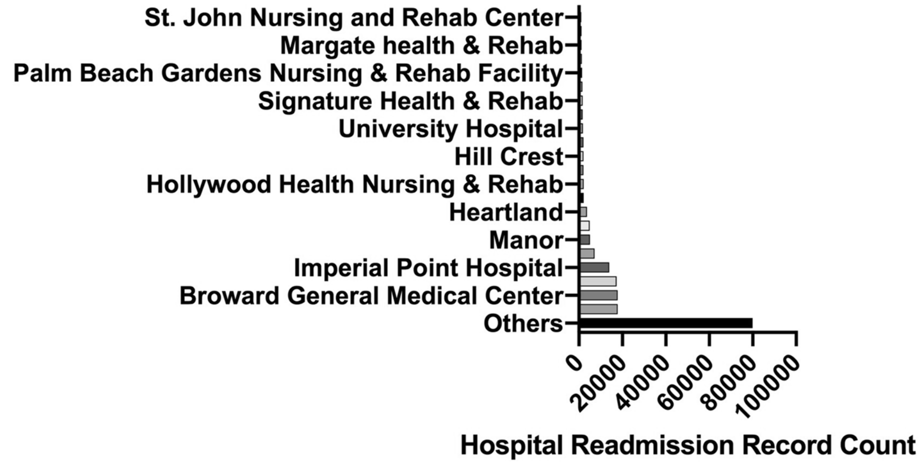

Fig. 2 Discharge summary count by hospital

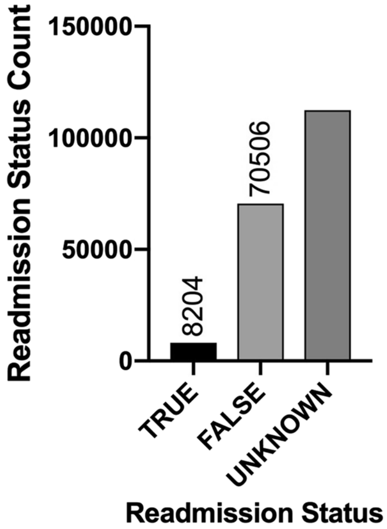

Fig. 3 Discharge summary count by readmission statuses

facilities with less than 1000 admissions are all grouped under others for data sanity and ease of use. The data is labelled with a readmit Boolean variable showing readmission status for all the admissions. Figure 3 shows the total number of records for readmission status of True, False, or Unknown.

The data was transferred from various care centers EHR systems to a HIPAA-compliant cloud services, where it was stored as MongoDB data dump. An open-source framework written in Python was used to preprocess the data. The preprocessing steps includes removal of special characters and removal of stop words from English language. The feature set was generated after applying Minimum and Maximum document frequency, i.e. Min *df* and Max *df*.

Table 1 Summary of extracted features and sample features per category

Category	Count	Sample features
Others	223	Lower, intact, history
Encounter reason	58	Headache, infection, lymphadenopathy
Location	44	Abdomen, kidney, chest
Medications	27	Creatinine, lasix, albuterol
Vitals	22	Vitamin, weight, glucose
health history	8	Alcohol, allergy, smoking
Procedures	5	Therapy, walker
Admissions	1	Clinic
Demographics	2	Female, male

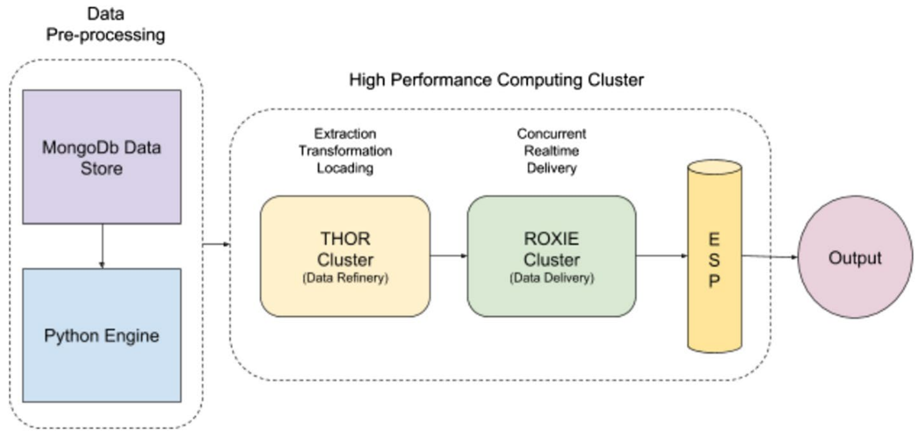

Fig. 4 Readmission analysis framework

This allowed removing of corpus specific stop words and removing words that appeared too infrequently. Different Min *df* and Max *df* values were used in this study to improve efficiency. A total of 389 distinct features were extracted from the database. The features extracted based on categories are shown in Table 1.

Framework used

As discussed in "Literature review" section, prior studies for readmission analysis have not addressed one of the very important real world application aspects of using Big Data. The readmission problem by its very nature a Big Data problem since the amount of data involved in developing effective and efficient prediction must be very high. If the dataset is smaller, the results are prone to be selective and limiting, and will provide skewed results which cannot be relied on. The inaccurate prediction may in turn cause negative clinical and financial outcomes for patients, family members, payers, or hospitals. In order to overcome this problem, the high performance computing cluster (HPCC) framework was used in this study.

The readmission analysis framework in this study has two main components: data pre-processing component and HPCC systems component. As shown in Fig. 4, the data pre-processing component has two main parts. The first part is a MongoDB Data

Store used to store all the discharge summaries and is used for some basic preprocessing like removing the generic discharge summary text which are not related to the actual discharge summary content. The second part of the pre-processing component is a Python Engine where the discharge summaries are classified into 389 features and 9 feature categories. The features are derived based on Min and Max document frequency, the Min *df* is 0.05 and Max *df* is 1.0. Stop words and any special characters are also removed after applying Document Frequency. Output from the data pre-processing component is then supplied to the HPCC component. HPCC systems component is an open source platform which provides a framework for analyzing data at any scale. The standard operating system is Linux on which the HPCC components can be deployed. The three main components of HPCC are Thor (Data Refinery cluster) which is a data storage and data refinery component, ROXIE (Rapid Online XML Inquiry Engine) which is a data delivery engine that also provides data warehousing capabilities, and ESP (Enterprise Services Platform) which enables end-users to access ROXIE queries via simple web protocols. The main advantage of using HPCC is its optimized distributed file system and massive scalability and performance. For example, THOR can process billions of records per second. HPCC programming is done in ECL (enterprise control language) which is declarative, modular, extensible, and is designed specifically for processing Big Data. Figure 5 shows the functional process flow used under this study.

Model training and evaluation

Initially, Logistic Regression and Nave Bayes methods were compared on the HPCC platform, and we noticed that Naive Bayes performed better in precision and recall. It also performed better from a Big Data perspective because of the ability to continue building on the current model when new data is available. This elasticity supports real time applications due to improved speed. Naive Bayes is well suited for our application wherein we will get new data and the model must adjust itself based on the new data. According to Liu et al. [43], Naive Bayes tends to perform better when the dataset increases which is a positive indicator in getting better accuracy. The study also suggests that Naive Bayes provides simple data fusion which allows algorithm to be flexible and elastic. The training time for Naive Bayes was also much faster than Logistic Regression.

Initially, the model was trained with *df* values between 0.2 and 1.0 with stop words included, 105 total features which was too low, and did not result in good accuracy.

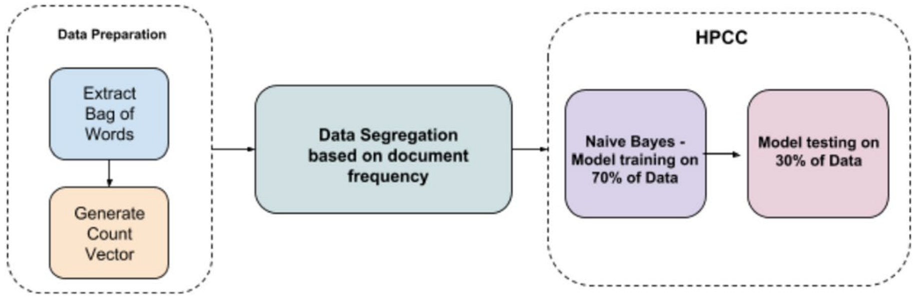
Data Preparation
Extract Bag of Words
Generate Count Vector
Data Segregation based on document frequency
HPCC
Naive Bayes - Model training on 70% of Data
Model testing on 30% of Data

Fig. 5 Functional process flow

Then, the *df* values were relaxed, with a Min *df* value of 0.1 and Max *df* of 1.0 with stop words included. This provided improved accuracy with 241 features. The data was again improved by using Min *df* of 0.05 and Max *df* of 1.0 with stop words excluded which resulted in highest accuracy with 389 features suggesting 0.05 *df* is an optimal cut-off for obtaining the highest accuracy.

Results and discussion

This study used Naive Bayes classification model to identify if a COPD patient would be readmitted. The dataset involved in this study and all similar studies is imbalanced because only a small subset of patients get re-admitted. Due to this, Accuracy is not a good measure for assessing model performance. This study used recall, precision and cluster time. Recall is the most important measure as recall shows the ability of a model to identify all the relevant cases within a dataset. Precision is also a very important factor in assessing model performance as it identifies the proportion of data points which are flagged as relevant, as actually being relevant. Table 2 shows the comparison of performance of Naive Bayes model with changes in Min *df* and Max *df* parameters. The best performing model in this study used Min *df* of 0.1, the Precision, Recall and Cluster time (which is overall Evaluation and Training time) is higher than the study performed by Jamei et al. [1] and LACE [16] which is an industry standard. Table 3 compares the performance measures of this study with these standards. Figures 6, 7, 8, 9 show comparison of AUC, Precision, Recall and Cluster Time based on different configurations and document frequency for Naive Bayes.

Precision

Table 2 shows that precision remains fairly stable for all values of Min *df* which means that precision does not change with change in incorporation of lower document frequency features except when the document frequency is made higher than 0.2 results in precision dropping to 77.2%.

Table 2 Comparison of various NB and DF configurations

Model	Min *df*	Max *df*	Stop words	# of features	AUC	Precision (%)	Recall (%)	Cluster time (s)
Naive Bayes	0.05	1	Y	389	0.544	90.65	50.15	824
Naive Bayes	0.1	1	Y	217	0.542	90.21	60.34	407
Naive Bayes	0.1	1	N	241	0.542	90.45	51	534
Naive Bayes	0.2	1	N	105	0.537	77.2	58	227

Table 3 Comparison of the performance of our models, model by Jamei et al. [1] and LACE [16]

Model	# of features	AUC	Precision (%)	Recall (%)	Time (s)
Naive Bayes (0.1 to 1)	217	0.542	90.21	60.34	407
2-layer neural network [1]	All	0.78	23	59	1040
LACE [16]	4	0.71	19	50	0

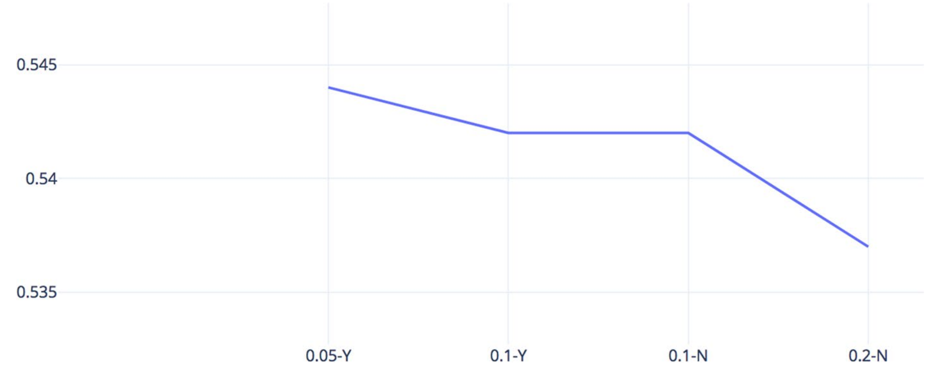

Fig. 6 Comparison of various NB and DF configurations—AUC

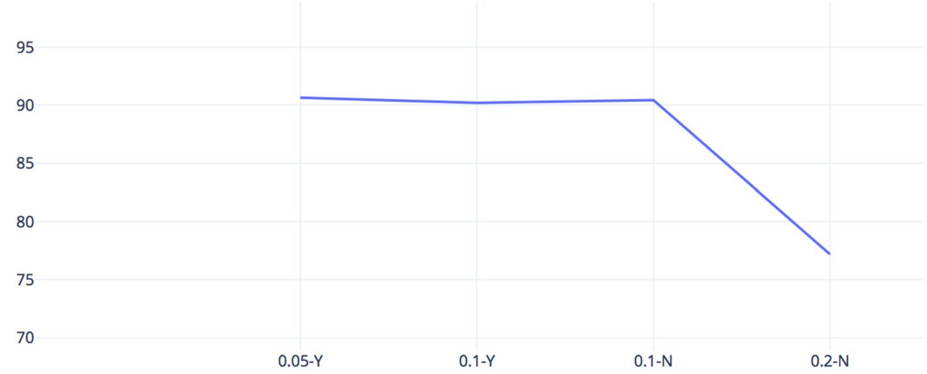

Fig. 7 Comparison of various NB and DF configurations—precision

Recall

Table 2 shows that Recall on the other hand changes with the incorporation of features with lower document frequency. Rare features show lower values of Recall whereas slightly higher values of document frequency show higher Recall. This is an indication that the rare words in a discharge summary do not help in identifying readmission probability in other patients as the rare words might be associated with a certain hospital chain or set of Physicians.

Cluster time

Cluster time is the overall time taken by a Cluster on the HPCC platform for training and evaluation. This shows that using HPCC platform has significantly reduced the overall training and evaluation time. The best performing model in this study was shown to be 60% faster than the model by Jamei et al. [1].

Conclusion

Various hospital readmissions studies have been conducted in different disease domains. These studies can be compared on the basis of the data type used, data size, disease conditions, algorithms and other features which would be valuable in solving the readmission problem. Based on the comparison it was understood that most of the researches were using high dimensionality data with high number of instances,

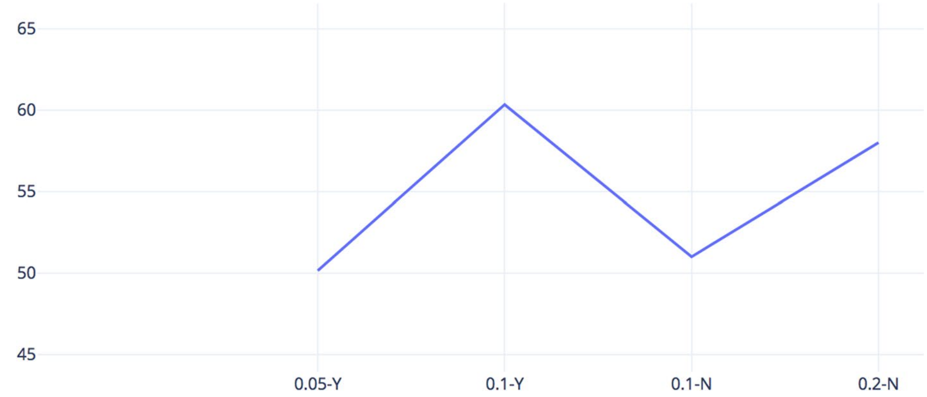

Fig. 8 Comparison of various NB and DF configurations—recall

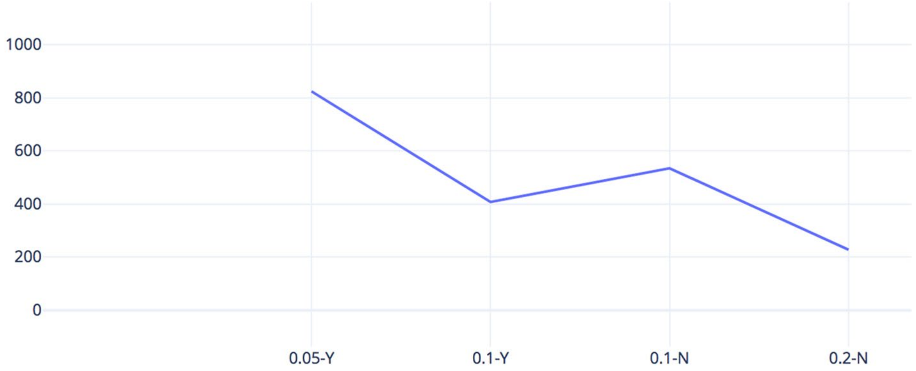

Fig. 9 Comparison of various NB and DF configurations—cluster time

but none of the researches used Big Data platforms. This study proposed a hospital readmission risk prediction framework which used the HPCC based Big Data platform, and showed that the overall training and testing time reduced by a significant margin while precision increased. This study shows that a real world implementation of a readmission risk prediction framework can effectively use a Big Data and parallel computing platform. The study also shows that the variety of discharge summaries written by multiple doctors for multiple Hospital Systems shows better performance as discussed in "Literature review" section. The results also shows that discharge summaries and clinical notes written by doctors at the time of discharge performs better than insurance claims data.

Abbreviations

HPCC: high performance computing cluster; COPD: chronic obstructive pulmonary disease; ROXIE: rapid online XML inquiry engine; ECL: enterprise control language; HRRP: Hospital Readmission Reduction Program; CMS: Centers for Medicare & Medicaid Services; ACA: Affordable Care Act; DRG: diagnose related group; IPPS: inpatient prospective payment system; R&D: research and development; CHF: congestive heart failure; THA/TKA: total hip arthroplasty/total knee arthroplasty; AMI: acute myocardial infarction; DAS: data analytics supercomputer; dF: document frequency; AUC: area under ROC curve; LACE: length of stay, acuity of admission, comorbidities and emergency department visits; ADHERE: acute decompensated heart failure registry; PPV: positive predictive value.

Authors' contributions

PJ carried out the conception, design, and implementation of this research as well as interpretation of results. AA, RB and CB made substantial contributions to the conception and design of this research as well as critically reviewing and interpreting results. PJ carried out the drafting of manuscript. AA, RB and CB critically reviewed the manuscript. All authors read and approved the final manuscript.

Author details

[1] Department of Computer & Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL, USA. [2] Department of IT and Operations Management, Florida Atlantic University, Boca Raton, FL, USA. [3] Department of Advanced Technology, Indian River State College, Fort Pierce, FL, USA.

Acknowledgements

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The data sets generated during and/or analyzed during the current study are not publicly available due to health care data privacy reasons and are not available and cannot be shared by corresponding author.

Funding

This research was supported in part by NSF Grants IIP-1444949 and IIP-1624497.

Publisher's Note

Received: 29 October 2018 Accepted: 8 March 2019
Published online: 16 March 2019

References

1. Mehdi J, Aleksandr N, Evrett W, Sylvia S, Eric L. Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. PLoS ONE. 2017;12(7):e0181173. https://doi.org/10.1371/journal.pone.0181173.
2. Goodman D, Fisher E, Chang C. The revolving door: A report on US Hospital Readmissions. Princeton: Robert Wood Johnson Foundation; 2013.
3. Readmissions Reduction Program (HRRP). CMS. https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html.
4. Christopher JLM, Alan DL. Alternative projections of mortality and disability by cause 1990–2020: global burden of disease study. Lancet. 1997;349:1498–504.
5. Surya PB, Michael W, et al. Results of a medicare bundled payments for care improvement initiative for chronic obstructive pulmonary disease readmissions. Ann Am Thorac Soc. 2017;14(5):643–8. https://doi.org/10.1513/annalsats.201610-775bc.
6. Tina S, Matthew MC, Marcelo CP, Tamara K. Understanding why patients with COPD get readmitted: a large national study to delineate the Medicare population for the readmissions penalty expansion. Chest. 2015;147:12191226.
7. Bernard F, Jayasree B. The rate and cost of hospital readmissions for preventable conditions. Med Care Res Rev. 2004;61(2):225240.
8. Mark M. Statement of executive director of the Medicare Payment Advisory Commission, before the Subcommittee on Health, Committee on Energy and Commerce. US House of Representatives. April 18, 2007.
9. Patricia H, Yves E, Isaline P. Validation of the potentially avoidable hospital readmission rate as a routine indicator of the quality of hospital care. Med Care. 2006;44(11):972981.
10. Sunil K, Frank L. Deficits in communication and information transfer between hospital based and primary care physicians. JAMA. 2007;297(8):831841.
11. Isaac S, Saeede A, Kai Y. A predictive analytics approach to reducing 30-day avoidable readmissions among patients with heart failure, acute myocardial in- farction, pneumonia, or COPD. Health Care Manag Sci. 2015;18(1):1934. https://doi.org/10.1007/s10729-014-9278-y.
12. Danning H, Simon CM, Anthony NK, Susan H. Mining high-dimensional administrative claims data to predict early hospital readmissions. J Am Med Inform Assoc. 2014;21(2):2729. https://doi.org/10.1136/amiajnl-2013-002151.
13. Joseph F, Jonathan M, Joseph L. A comparison of models for predicting early hospital readmissions. J Biomed Inform. 2015;56:22938. https://doi.org/10.1016/j.jbi.2015.05.016.
14. Ruben A, Billy JM, Ying PT, Mark HD, Clark CA, Song Z, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Med Care. 2010;48(11):9818. https://doi.org/10.1097/MLR.0b013e3181ef60d9.
15. Mohsen B, Mark B, Michael G, Karen MM, George R, Mark SS, Eric H. Data-driven decisions for reducing readmissions for heart failure: general methodology and case study. PLoS ONE. 2014;9(10):e109264. https://doi.org/10.1371/journal.pone.0109264.
16. Carl VW, Irfan AD, Chaim B, Edward E, Ian GS, Kelly Z, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. Can Med Assoc J. 2010;182(6):5517. https://doi.org/10.1503/cmaj.091117.

17. Devan K, Honora E, Amanda S, David K, Cecelia T, Michele F, Sunil K. Risk prediction models for hospital readmission: a systematic review. JAMA. 2011;306(15):168898. https://doi.org/10.1001/jama.2011.1515.
18. David WB, Suchi S, Lucila O, Anand S, Gabriel E. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff. 2014;33(7):1123. https://doi.org/10.1377/hlthaff.2014.0041.
19. Stephen FJ, Mark VW, Eric AC. Rehospitalizations among patients in the Medicare fee-for-service program. N Engl J Med. 2009;360(14):1418.
20. Carolyn MC. Commentary: reducing hospital readmissions: aligning financial and quality incentives. Am J Med Qual. 2012;27(5):441.
21. Robert PK, Eli YA. Hospital readmissions and the Affordable Care Act: paying for coordinated quality care. JAMA. 2011;306(16):1794.
22. Xiaoqian J, Aziz AB, Robert E, Kim J, Lucila O. A patient-driven adaptive prediction technique to improve personalized risk estimation for clinical decision support. J Am Med Inform Assoc. 2012;19(e1):e36.
23. Suchi S, Anand KR, Jeffrey G, Daphne K, Anna AP. Integration of early physiological responses predicts later illness severity in preterm infants. Sci Transl Med. 2010;2(48):48ra65.
24. Tobias F, Cornelia UK, Dominik O, Frank P. Patterns of Multimorbidity in primary care patients at high risk of future hospitalization. Popul Health Manag. 2012;15(2):119.
25. Seref S, Duygu S. Big data: a review. In: 2013 international conference on collaboration technologies and systems (CTS), San Diego. 2013, p. 42–7. https://doi.org/10.1109/cts.2013.6567202.
26. Hengyu C, Hang Z, Yutao L, Guijie L. Research on application of healthcare data in big data era. In: 2018 international conference on robots and intelligent system (ICRIS). Changsha. 2018, p. 377–9. https://doi.org/10.1109/icris.2018.00100.
27. Matthew H, Taghi MK, Richard AB. Big Data fraud detection using multiple medicare data sources. J Big Data. 2018;5:29. https://doi.org/10.1186/s40537-018-0138-3.
28. Christopher B, Ankur A. A framework for the estimation and reduction of hospital readmission penalties using predictive analytics. J Big Data. 2017;4:37. https://doi.org/10.1186/s40537-017-0098-z.
29. Ankur A, Christopher B, Ravi SB, Vinaya R. Multi-method approach to wellness predictive modeling. J Big Data. 2016;3:15. https://doi.org/10.1186/s40537-016-0049-0.
30. Christopher B, Ankur A, Ravi B, Xingquan Z. A cost sensitive approach to pre-dicting 30-day hospital readmission in COPD patients. In: 2017 IEEE EMBS international conference on biomedical and health informatics (BHI). Orlando. https://doi.org/10.1109/bhi.2017.7897269.
31. Christopher B, Ankur A, Ravi B, Xingquan Z. Latent topic ensemble learning for hospital readmission cost reduction. In: 2017 international joint conference on neural networks (IJCNN). Anchorage. https://doi.org/10.1109/ijcnn.2017.7966439.
32. Anthony M, David B, Gavin H, Arjuna C, Borko F. The HPCC/ECL platform for big data big data technologies and applications. Cham: Springer; 2016.
33. Borko F, Flavio V. Introduction to big data. Big data technologies and applications. Berlin: Springer; 2016. p. 311.
34. Lili X, Edin M, Amy A. ECL-watch: a big data application performance tuning tool in the HPCC systems platform. In: 2017 IEEE international conference on big data (Big Data), Boston. 2017, p. 2941–50. https://doi.org/10.1109/bigdata.2017.8258263.
35. David B. Aggregated data analysis in HPCC systems. Big data technologies and applications. Cham: Springer; 2016.
36. Jeffrey D, Sanjay G. MapReduce: simplified data processing on large clusters. Commun ACM. 2008. https://doi.org/10.1145/1327452.1327492.
37. Michael P, Linh N, Flavio V, Amy A. Managing the academic data lifecycle: a case study of HPCC systems. In: 2014 IEEE International Conference on big data. 2014, p. 22–30.
38. Maryam M, Taghi MK, Flavio V, John H. Large-scale distributed L-BFGS. J Big Data. 2017;4:22. https://doi.org/10.1186/s40537-017-0084-5.
39. LexisNexis risk solutions. HPCC systems for cyber security analytics. September 2012.
40. Borko F, Flavio V. Social network analytics: hidden and complex fraud schemes. Big data technologies and applications. Cham: Springer; 2016.
41. Flavio V, Mauricio R. HPCC systems for cyber security analytics. big data technologies and applications. Cham: Springer; 2016.
42. LexisNexis risk solutions. The rise of organized crime in health care: social network analytics uncover hidden and complex fraud schemes. December 2011.
43. Liu B, Blasch E, Chen Y, Shen D, Chen G. Scalable sentiment classification for Big Data analysis using Nave Bayes Classifier. In: 2013 IEEE international conference on big data. Silicon Valley. 2013, p. 99–104. https://doi.org/10.1109/bigdata.2013.6691740.