USING DATA MINING TO PREDICT HOSPITAL ADMISSIONS FROM THE EMERGENCY DEPARTMENT

Hemant Reddy Palavali

Dept. of Computer Science and Engineering, Jawaharlal Nehru Technological University, Hyderabad, India

ABSTRACT:

Crowding within emergency departments (EDs) can have significant negative consequences for patients. EDs therefore need to explore the use of innovative methods to improve patient flow and prevent overcrowding. One potential method is the use of data mining using machine learning techniques to predict ED admissions. This paper uses routinely collected administrative data (120 600 records) from two major acute hospitals in Northern Ireland to compare contrasting machine learning algorithms in predicting the risk of admission from the ED. We use three algorithms to build the predictive models: 1) logistic regression; 2) decision trees; and 3) gradient boosted machines (GBM). The GBM performed better (accuracy = 80.31%, AUC-ROC = 0.859) than the decision tree (accuracy = 80.06%, AUC-ROC = 0.824) and the logistic regression model (accuracy = 79.94%, AUC-ROC = 0.849). Drawing on logistic regression, we identify several factors related to hospital admissions, including hospital site, age, arrival mode, triage category, care group, previous admission in the past month, and previous admission in the past year. This paper highlights the potential utility of three common machine learning algorithms in predicting patient admissions. Practical implementation of the models developed in this paper in decision support tools would provide a snapshot of predicted admissions from the ED at a given time, allowing for advance resource planning and the avoidance bottlenecks in patient flow, as well as comparison of predicted and actual admission rates. When interpretability is a key consideration, EDs should consider adopting logistic regression models, although GBM's will be useful where accuracy is paramount.

KEYWORDS: Data Mining, Emergency Department, Hospitals, Machine Learning, Predictive Models

I. INTRODUCTION

In recent years we have witnessed a dramatic increase of electronic health data, including extensive Electronic Medical Records (EMR) recording patient conditions, diagnostic tests, labs, imaging exams, genomics, treatments, outcomes, claims, financial records, clinical guidelines and best practices etc[1]. Healthcare professionals are now increasingly asking the question: what can we do with this wealth of data? How can we perform meaningful analytics on such data to derive insights to improve quality of care and reduce cost? Healthcare Analytics needs to cover the whole spectrum including both Knowledge Driven Analytics and Data Driven Analytics[2]. Knowledge driven approaches operate on knowledge repositories that include scientific literature, published clinical trial results, medical journals, textbooks, as well as clinical practice guidelines. Traditionally the gold standard of evidence in healthcare has been produced through the randomized controlled trial process.

While most emergency department (ED) visits end in discharge, EDs represent the largest source of hospital admissions. In the ED, patients are first sorted by acuity in order to prioritize individuals requiring urgent medical intervention. This sorting process, called "triage", is typically performed by a member of the nursing staff based on the patient's demographics, chief complaint, and vital signs. Subsequently, the patient is seen by a medical provider who creates the initial care plan and ultimately recommends a disposition, which this study limits to hospital admission or discharge [3,4].

Prediction models in medicine seek to improve patient care and increase logistical efficiency. For example, prediction models for sepsis or acute coronary syndrome are designed to alert providers of potentially life-threatening conditions, while models for hospital utilization or patient-flow enable resource optimization on a systems level. Early identification of ED patients who are likely to require admission may enable better optimization of hospital resources through improved understanding of ED patient mixtures. It is increasingly understood that ED crowding is correlated with poorer patient outcomes. Notification of administrators [5] and inpatient teams regarding potential admissions may help alleviate this problem. From the perspective of patient care in the ED setting, a patient's likelihood of admission may serve as a proxy for acuity, which is used in a number of downstream decisions such as bed placement and the need for emergency intervention [6].

Numerous prior studies have sought to predict hospital admission at the time of ED triage. Most models only

include information collected at triage such as demographics, vital signs, chief complaint, nursing notes, and early diagnostics, while some models include additional features such as hospital usage statistics and past medical history [7]. A few models built on triage information have been formalized into clinical decision rules such as the Sydney Triage to Admission Risk Tool and the Glasgow Admission Prediction Score [8]. Notably, a progressive modeling approach that uses information available at later time-points, such as lab tests ordered, medications given, and diagnoses entered by the ED provider during the patient's current visit, has been able to achieve high predictive power and indicates the utility of these features. We hypothesized that extracting such features from a patient's previous ED visits would lead to a robust model for predicting admission at the time of triage. Prior models that incorporate past medical history [9] utilize simplified chronic disease categories such as heart disease or diabetes while leaving out rich historical information accessible from the electronic health record (EHR) such as outpatient medications and historical labs and vitals, all of which are routinely reviewed by providers when evaluating a patient. In this work showed that using all elements of the electronic health record can robustly predict in-patient outcomes, a prediction model for admission built on comprehensive elements of patient history may improve on prior models [10].

It focuses on developing and applying machine learning and data mining tools to an array of different challenging problems from clinical genomic analysis, through designing clinical decision support systems. There are two general categories of algorithms: unsupervised and supervised. Unsupervised machine learning algorithms are typically used to group large amounts of data. Unsupervised algorithms can be used to generate hypotheses, and thus, often precede use of a supervised algorithm. Supervised machine-learning algorithms start out with a hypothesis and categories that are set out in advance. These results are then used to make predictions based on out-of-sample data for which the outcome of interest is not known.

II. LITERATURE SURVEY

Byron Graham. [1] developed a prediction model in which machine learning techniques such as Logistic Regression, Decision Tree and Gradient Boosted Machine were used. The most important predictors in their model were age, arrival mode, triage category, care group, admission in past-month, past-year. In which the gradient boosted machine outperforms and focus on avoiding the bottleneck in patient flow.

Jacinta Lucke. [2] and team has designed the predictive model by considering age as main attribute, where the age is categorized in two categories below 70 years and above 70 years. They observed that the category of people below 70 years was less admitted when compared with the category of people above 70 years. Younger patient group had higher accuracy while the older patient group had high risk of getting admitted to hospital. The decision of prediction was based on the attributes such as age, sex, triage category, mode of arrival, chief complaint, ED revisits, etc.

Xingyu Zhang [3] in their predictive model, they have used logistic regression and multilayer neural network. These methods were implemented using natural language processing and without using natural language processing. The accuracy of model with natural language processing is more than the model without natural language processing.

Boukenze. [4] with his team created a model using decision tree C4.5 for predicting admissions which overall gave a good accuracy and less execution time. The author has used the prediction model for predicting a particular disease that is chronic kidney disease.

Dinh and his team [5], developed a model which uses multivariable logistic regression for prediction. For the prediction the two main attributes were demographics and triage process, which helped to increase the accuracy.

Davood. [6] developed a model for reducing emergency department boarding using the logistic regression and neural network, where a set of thumb rules were developed to predict the hospital admissions. The prediction model used as decision support tool and helped to reduce emergency department boarding. The set of thumb rules were found by examining the importance of eight demographic and clinical factors such as encounter reason, age, radiology exam type, etc. of the emergency department patient's admission. Xie. [7] and his teams model consist of coxian phase type distribution (PH Model) and logistic regression where the PH model has out performs than logistic regression.

Peck and his teams [8] created a model for predicting the inpatient for same-day to improve patient flow. The model uses Naive Bayes and linear regression with logit link function, the result of the model was accurate even though it had less number of independent variables. Sun. [9] and his team uses logistic regression for creating the model with the help of triage process which plays an important role for early prediction of hospital admission

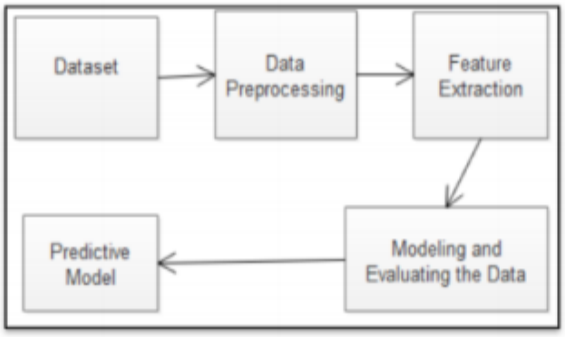
The factors which were considered for prediction are age, sex, emergency visit in preceding three months, arrival mode, patient acuity category, coexisting chronic diseases. Jones. [10] with his team developed a predictive model for forecasting the daily patient volumes in the emergency department. The model uses regression which is actually a time series regression and exponential smoothing where

time series regression performs better than linear regression.

III. SYSTEM ANALYSIS

SYSTEM ARCHITECTURE

The system architecture consists of five steps:



1) Dataset: A hospital dataset is taken for the further processing. This is the raw data in the comma separated value (csv) format. The dataset consists of 10 attributes such as ED-level, acuity, etc

2) Data Preprocessing: The second step is data preprocessing in which all the null values, missing values are removed. Removal of extra value is also done. Format the attributes in correct format.

3) Feature Extraction: In this step, a particular number of features are extracted for the model. Such features are selected which are important and which help in predicting

4) From the 10 attributes, main attributes are selected. Here five main attributes are selected.

a) OSHPD-ID: A unique 10-digit number assigned to each patient.

b)ED-LEVEL: Hospital services providing immediate initial evaluation and treatment to patients on a 24hrs basis.

c) EMSA-TRAUMA-LEVEL: Emergency Medical

d) Services Agency trauma center designation level.

e) ACUITY: Emergency Department type of visit.

f) ADMISSION-FROM-ED: Total Emergency Department visits by type, resulting in an inpatient admission.

5) Modeling the data: The complete dataset is divided into two parts, training and testing. In this step the training dataset is used. Using different machine learning techniques, the model is trained. The trained model obtained in the previous step is now used to evaluate. For evaluating, the testing dataset is used

6) Predictive Model: Now after the number of times training and evaluating the model, it is ready for the prediction purpose where external data is given as input.

EXISTING SYSTEM In Existing system, the single data mining technique is used to predict hospital admissions from the emergency department. There is no previous research that identifies which data mining technique can provide more reliable accuracy in identifying suitable treatment for hospital admissions from the emergency department. Practical use of hospital database systems and knowledge discovery is difficult in hospital admissions from the emergency department [17].

DISADVANTAGES

1. Hospitals do not provide the same quality of service even though they provide the same type of service.

2. There is no previous research that identifies which data mining technique can provide more reliable accuracy in identifying suitable solution to predict hospital admissions from the emergency department.

3. It takes more time consumption for practical use of hospital database systems.

PROPOSED SYSTEM

In Proposed System, we are applying data mining techniques (Hybrid) in identifying suitable solution to predict hospital admissions from the emergency department. Apply single data mining techniques to predict hospital admissions from the emergency department is benchmark dataset to establish baseline accuracy for each single data mining technique to predict hospital admissions from the emergency department. Apply the same single data mining techniques used in hospital admission to predict hospital admissions from the emergency department dataset to investigate if single data mining technique can achieve equivalent (or better) results in suitable solution identifying to predict hospital admissions from the emergency department. Apply hybrid data mining techniques to predict hospital admissions from the emergency department benchmark dataset to establish baseline accuracy for each hybrid data mining technique in the to predict hospital admissions from the emergency department. Apply the same hybrid data mining techniques used in hospital admission to predict hospital admissions from the emergency department dataset to investigate if hybrid data mining techniques can achieve equivalent (or better) results in identifying suitable solution identifying to predict hospital admissions from the emergency department [18].

ADVANTAGES

1. By applying data mining techniques to help emergency department in hospital to predict hospital admissions from the emergency department.

2. Hybrid data mining techniques are used for selecting the suitable to predict hospital admissions from the emergency department.

3. Time consumption is less.

4. High Performance and Accuracy

In this segment different existing methods have been talked about. Distributed storage is viewed as an arrangement of dispersed server farms that for the most part Utilize virtualization, innovation and supplies interface for information stockpiling.

Machine Learning Techniques in Healthcare

Machine learning (ML) provides methods, techniques and tools that can help solving diagnostic and prognostics problems in a variety of medical domains. ML is being used for the analysis of the importance of clinical parameters and of their combinations for prognosis, e.g. prediction of disease progression, for the extraction of medical knowledge for outcomes research, for therapy planning and support, and for overall patient management.

Types of Machine Learning Algorithms

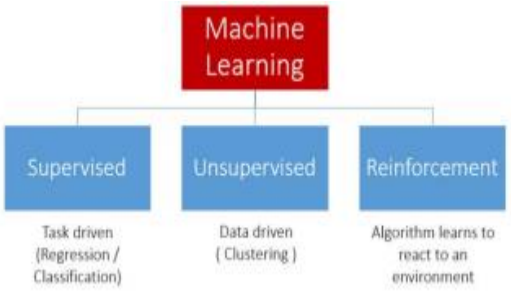
Four different types of machine learning algorithms are available that can be organized into taxonomy based on the desired outcome of the algorithm or the type of input available for training the machine. Thompson noted, "The terminology used in machine learning is different than that used for statistics. For example, in machine learning, a target is called a label, while in statistics it's called a dependent variable." [23] The key types of machine learning include:

Supervised learning

Unsupervised learning

Semi supervised learning

Reinforcement learning



Supervised learning is a type of machine learning that uses a known dataset (called the training dataset) to make predictions. The training dataset includes input data and labelled response values. Supervised machine learning techniques are more suitable for medical data classification. Unsupervised learning is a type of machine learning used to draw inferences from datasets consisting of input data without labelled responses.

Applications of Machine Learning Techniques in Health Care

Machine learning algorithms are effective in recognizing complex patterns within rich and massive data. This capability is particularly well-suited to medical applications, especially those that depend on complex proteomic and genomic measurements. As a result, machine learning is frequently used in various disease diagnosis and detection. In clinical applications machine learning algorithms can produce better decisions about treatment plans for patients by means of providing effective healthcare system

IV. MACHINE LEARNING ALGORITHMS AND PERFORMANCE

In this model three machine learning algorithms are used for training purpose: (1) Gradient Boosted Machine, (2) Random Forest and (3) Decision Tree. Boosting is a class of ensemble learning techniques for classification problems. It aims to build a set of weak learners to create one strong learner. The gradient boosted machine is that algorithm which is a tree based ensemble technique. GBM creates multiple weakly associated decision trees that are combined to get the final prediction. It is also known as boosting model. The second algorithm is the Random forest. This algorithm also uses an ensemble learning approach for classification while training process by creating number of decision trees. The next algorithm is the decision tree which is specifically recursive partitioning. The algorithm splits the data at each node based on the variable that separates the data unless an optimal model is not obtained [1].

Using RPART, CARET packages the implementation of the above algorithm is done. As decision tree works on single tree and the random forest and gradient boosted

machine works on ensemble of trees this packages are helpful to implement. The CARET package was used to train and tune the machine learning algorithms. This library provides a consistent framework to train and tune models. The performances of the machine learning algorithms are evaluated by the range of measures such as Accuracy, Cohens Kappa, Sensitivity and Specificity.

V. CONCLUSION AND FUTURE WORK

The overall study involved a survey of different methods used for the prediction model of hospital admission. Along with this study it also compares three different machine learning algorithms namely, decision tree, random forest and gradient boosted machine which are used for predicting the hospital admission from the emergency department. Overall the random forest performs better when compared to the decision tree and gradient boosted machine. Implementation of these models could help the hospital decision makers for planning and managing the hospital resources based on the patient flow. This would help reducing the emergency department crowding.

In future, different learning and machine learning algorithms can be used to implement the model. Even ensemble of different algorithms can also be done. Different demographics as predictor can be taken into consideration

REFERENCES

[1]. Wright SP, Verouhis D, Gamble G, Swedberg K, Sharpe N, Doughty RN. Factors influencing the length of hospital stay of patients with heart failure. Eur J Heart Fail. 2003; 5(2):201–209.

[2]. Gomez V, Abasolo JE. Using data mining to describe long hospital stays. Paradigma. 2009; 3(1):1–10.

[3]. Lim A, Tongkumchum P. Methods for analyzing hospital length of stay with application to inpatients dying in Southern Thailand. Glob J Health Sci. 2009; 1(1):27–38.
[4]. Chang KC, Tseng MC, Weng HH, Lin YH, Liou CW, Tan TY. Prediction of length of stay of first-ever ischemic stroke. Stroke. 2002; 33(11):2670–2674.

[5]. Jiang X, Qu X, Davis L. Using data mining to analyze patient discharge data for an urban hospital. In : Proceedings of the 2010 International Conference on Data Mining; 2010 Jul 12-15; Las Vegas, NV. p. 139–144.

[6]. Isken MW, Rajagopalan B. Data mining to support simulation modeling of patient flow in hospitals. J Med Syst. 2002; 26(2):179– 197.

[7]. Walczak S, Scorpio RJ, Pofahl WE. In : Cook DJ, editor. Predicting hospital length of stay with neural networks. Proceedings of the Eleventh International FLAIRS Conference; 1998 May 18-20; Sanibel Island, FL. Menlo Park, CA: AAAI Press;1998. p. 333–337.

[8]. Rowan M, Ryan T, Hegarty F, O'Hare N. The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors. Artif Intell Med. 2007; 40(3):211–221.

[9]. Robinson GH, Davis LE, Leifer RP. Prediction of hospital length of stay. Health Serv Res. 1966; 1(3):287–300.

[10]. Arab M, Zarei A, Rahimi A, Rezaiean F, Akbari F. Analysis of factors affecting length of stay in public hospitals in Lorestan Province, Iran. Hakim Res J. 2010; 12(4):27–32.

[11]. Blais MA, Matthews J, Lipkis-Orlando R, Lechner E, Jacobo M, Lincoln R, et al. Predicting length of stay on an acute care medical psychiatric inpatient service. Adm Policy Ment Health. 2003; 31(1):15–29.

[12]. Tu JV, Guerriere MR. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. Proc Annu Symp Comput Appl Med Care. 1992; 31(1):666–672