International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET231012

Predictive Analysis of Heterogenous Data for Hospital Readmission

V R Reji Raj, Mr. Rasheed Ahammed Azad .V
Department of Computer Science, Govt. Engineering College, Idukki, India

Article Info

Publication Issue :
Volume 10, Issue 1
January-February-2023
Page Number : 106-112

Article History
Accepted : 05 Jan 2023
Published: 19 Jan 2023

ABSTRACT

Hospital readmission is a high priority health care quality measure. Diabetes patient readmission rate is increasing to such an extent that it becomes one of the major concerns for many hospitals. Many studies are conducted for finding the possible causes and risks of diabetes patient's readmission. Reducing readmission rates of diabetes patients reduce health care costs. Here the relationship between diabetes and the various patient attributes are examined. Different prediction models were developed to predict the risk of readmission within 30 days among hospitalized patients with diabetes. The dataset used here contains more than 1 lakh observations and 56 features. They include a set of numerical attributes such as number of outpatient visits, number of emergency visits and time spent in hospital etc and a set of categorical data such as what type of admission the encounter faced , sets of drugs that the patient took etc. In this study we presented a scheme to identify high-risk patients and evaluated different machine learning algorithms. Results indicate that Adaboost with hyperparameter tuning is optimal for this task The results from the study help health care providers to improve diabetic care.

Keywords: Readmission Prediction, Data mining, Hyperparameter, Adaboost, Hyperparameter tuning, Receiver Operating Characteristics, Area under the ROC Curve.

I. INTRODUCTION

Diabetes is one of the most widely spread life threatening chronic disease that is associated the variation in insulin levels in the blood glucose level. In healthcare hospital readmission is considered an effective measure of care. Hospital readmission is the time that a patient takes, before getting back to the hospital. Patients with high risk for readmission need to be identified at the time of being discharged from the hospital, to facilitate improved treatment.

Readmission of patients within 30 days of being discharged has been a widely used metric for studying readmissions. However, it's a fact that a significant number of diabetic patients are readmitted after 30 days from discharge. Readmission is considered as a quality measure of hospital performance as well as a mean to reduce healthcare costs. Current practices used are patient visits a doctor and he will assess the patient and decide what the appropriate care plan for that person.

Machine learning plays a vital role in many predicting tasks. Hence, hospital readmission prediction using machine learning sounds a worth implementing approach. These models are more complex, but may be able to create more accurate risk predictions that should lead to improved diabetic patient outcomes. By using machine learning on a wide feature set for prediction improve the accuracy of readmission risk. Readmission can be predicted by using different data mining techniques. This includes decision trees, logistic Regression etc. These approaches help the health care.

The dataset used here contains more than one lakh admissions each containing 50 features. Here we examine diabetes patient readmission rate using classification. Categorical data such as admission source, admission type, discharge position and the set of drugs etc also analyzed which has any impact on the readmission rate or not. The dataset contains data systematically collected from participating institutions, electronic medical records and includes encounter data such as emergency, outpatient, and inpatient, provider specialty, age, sex, and race, diagnoses and in-hospital procedures documented by ICD-9-CM codes, laboratory data, pharmacy data, in-hospital mortality, and hospital characteristics.

II. RELATED WORKS

Many researches were conducted in this field. In the work Decision Support in Heart Disease Prediction System using Naive Bayes, data mining technique is used for uncovering relations that connect variables in the database. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It is implemented as web based questionnaire application. This method can also be used as a training tool to train nurses and medical students to diagnose patients with heart disease.

In Decision Tree Discovery for the Diagnosis of Type II Diabetes, decision tree method was used to predict patients with developing diabetes. The dataset used was the Pima Indians Diabetes Data Set, which collects the information of patients with and without developing diabetes. The study was conducted in two phases. The first phase is data preprocessing including attribute identification and selection, handling missing values, and numerical discretization. The second phase is a diabetes prediction model construction using the decision tree method. Weka software was used throughout all the phases of this study.

In another work Classification of Heart Disease Using KNN and Genetic Algorithm, KNN is used for pattern recognition.KNN is a straight forward classifier, where samples are classified based on the class of their nearest neighbor. If the data set contains redundant and irrelevant attributes, classification may produce less accurate result. In this paper a new algorithm which combines KNN with genetic algorithm was used for effective classification. Genetic algorithms perform global search in complex large and multimodal landscapes and provide optimal solution. Experimental results shows that the algorithm enhance the accuracy in diagnosis of heart disease.

Existing models developed to predict readmissions for pneumonia lack the required accuracy in prediction. In the study, Design of a Clinical Decision Support Model for Predicting Pneumonia, they determined the risk factors to predict readmission. A clinical decision support system (CDSS) was designed to predict readmission within 30 days after discharge. The selected features are then applied using the RBF-SVM. About 16.2% patients were readmitted. Variables such as age, gender, number of medication, length of admission and total admission cost were observed to be significant.

DESIGN

1. Overview of the Method

An overview of the method is shown in Fig.1.

Fig. 1. Overview of the design

Data understanding means we have to know full details about the data, how it is related to target variable etc.

In **preprocessing** steps the multivariate data sets are analyzed. The target set is then cleaned. Data cleaning removes features containing noise and those with missing data.

Data mining involves six common classes of tasks.

- **Anomaly detection**: process of identifying unusual data records that require further investigation.
- **Association rule learning:** Process of searching for relationships between variables.
- **Clustering**: Task of discovering groups and structures in the data that are in some way or another similar without using known structures in the data.
- **Classification**: Task of generalizing known structure to apply to new data.
- **Regression**: attempts to find a function which models the data with the least error that is, for estimating the relationships among data or datasets.
- **Summarization:** provides a more compact representation of the data set, including visualization and report generation.

In the **evaluation phase**, algorithm was applied on the test dataset which was not trained. The learned patterns are then applied to this test dataset, and the output is compared to the desired output. Once trained, the learned patterns would be applied to the test set. If we do not meet the desired standards, re-evaluate and change the pre-processing and data mining steps. If the desired standards meet, then interpret the learned patterns and turn them into knowledge.

III. IMPLEMENTATION

Implementation literally means to put into effect what to carry out. The system implementation phase deals with translation of the design specification into the source code. The ultimate goal of implementation is to write the source code and the internal documentations. This eases debugging, testing and modifications. System flow starts by simply running on packages, sample output etc. is a part of implementation.

By implementing the project, the project should meet the following goals:

- Clarity and simplicity of code.
- Accuracy of result.
- Minimization of response code.
- Minimization of amount of memory used.

1. System architecture

System architecture is shown in Fig.2. The method for this study involves the following data mining tasks.

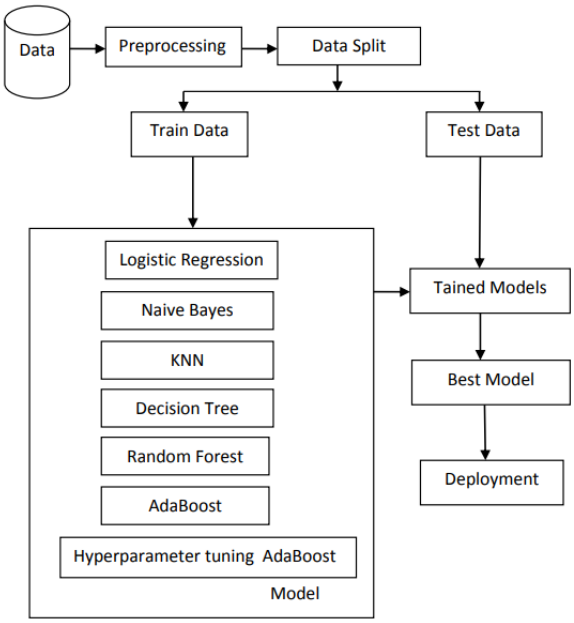

Fig 2: System Architecture

1.1 Dataset description

Finding a good dataset is one of the first challenges. Here used is a publicly available dataset from UCI repository containing diabetes patient encounter data for 130 US hospitals (1999– 2008) containing 101,766 observations over 10 years. The dataset has over 50 features including patient characteristics, conditions, tests and 23 medications.

1.2 Preprocessing

Before actual modeling, some wrangling with the data is always needed. We applied three types of methods here:

- **Cleaning** tasks such as dropping bad data, dealing with missing values etc.
- **Modification** of existing features e.g. normalization.
- **Creation or derivation** of new features, usually from existing ones.

The dataset contained up to three diagnoses for a given patient- primary, secondary and additional. We collapsed these diagnosis codes into 9 disease categories. These 9 categories include Circulatory, Respiratory, Digestive, Diabetes, Injury, Musculoskeletal, Genitourinary, Neoplasm, and Others. We use only the primary diagnosis in our model.

Variables such as number of inpatient, emergency room visits and outpatient visits measures how much hospital/clinic services a person has used in the past year. We added these three to create a new variable called service utilization. The dataset contains 23 features for 23 drugs, whether a change in that medication was made or not during the current hospital stay of patient is to be determined. We decided to count how many changes were made in total for each patient, and declared that a new feature.

Another possibly related factor could be the total number of medications used by the patient. So we created another feature by counting the medications used during the encounter. Just like diagnoses, there were quite a few categories for admission source, admission type and discharge disposition. We collapsed these variables into fewer categories where it made sense. For example, admission types 1, 2 and 7 correspond to Emergency, Urgent Care and Trauma, and thus were combined into a single category.

The original dataset used string values for gender, race, medication change, and each of the 23 drugs used. To better fit those variables into our model, we interpret the variables to numeric binary variables to reflect their nature. The outcome we are looking at is whether the patient gets readmitted to the hospital. The variable actually has < 30, > 30 and No Readmission categories.

To reduce the problem to a binary classification, readmission after 30 days and no readmission are combined into a single category. The dataset only gives us age as 10 year categories, so we don't know the exact age of each patient. To do that, we assume that age of the patient on average lies at the midpoint of the age

category. Some patients in the dataset had more than one encounter. We decided to use first encounters of patients with multiple encounters. This resulted in dataset being reduced to about 70,000 encounters.

2. Data splitting

The final dataset consisted of 120,600 observations of which 10.8% had missing data, leaving 107,545 cases for building the models. To enable validation of the model, random stratified sampling was used to split the data into training (70% of cases) and test (30% of cases) datasets.

3. Classification

Prior to training the classification algorithms, split the dataset into two distinct sets - the training and the test set. The training and test set consisted of 70% and 30% of the data. The performance of all algorithms was evaluated on the test set. The models that implemented include:

1. **Logistic regression**: With the starting assumption that the impact of factors and their interactions can be modeled as a log likelihood of outcome, logistic regression can help us understand the relative impact and statistical significance of each factor on the probability of readmission.

2. **Naive Bayes**: Naive Bayes algorithm is a probabilistic model for classification. It assumes that given the class, features are statistically independent of each other. Naive bayes uses time-sequence information of what came before (prior) and what came after (posterior) the variable being predicted.

3. **K-nearest Neighbors**: While K-nearest neighbors provide decent predictions, they cannot help in deciding the features that contribute to this decision the most, since features are weighted equally (assuming we normalize them) based on simply their contribution to the proximity/distance function.

4. **Decision Trees:** By iteratively and hierarchically observing the level of certainty of predicting whether someone would be readmitted or not, we find the relative importance of different factors using a more human-like decision making strategy in establishing this determination.

5. **Random Forests**: By considering more than one decision tree and then doing a majority voting, random forests helped in being more robust predictive representations than trees as in the previous case. Each decision tree acts as a weak classifier and pooling the responses from multiple decision trees leads to a strong classifier.

6. **Adaboost**: In Adaboost a strong classifier is build by sequentially combining a set of weak classifiers. At first iteration, a single classifier is learnt to minimize the classification error. At each successive iteration, a new classifier is learnt which minimize the errors from previous iterations. Here used decision trees as weak classifiers.

7. **Hyperparameter Tuning: It** includes the design decisions that are made when we set up the machine learning model. Here we optimize the hyper parameters for AdaBoost classifier. One technique for hyperparameter tuning is Grid search in which we test all possible combinations over a grid of values. We can see that the hyperparameter tuning improved the models, but not by much. This is most likely due to the fact that we have a high bias situation. More improvement would be expected if we had high variance.

IV. Result Analysis

A system for predicting high risk patients useful when a large fraction of patients at high-risk, are correctly identified (i.e. high recall) without raising a large number of false alarms (i.e. high precision). Receiver Operating Characteristics (ROC) is one of the best metrics to evaluate classification models. ROC is a probability curve and AUC represents degree or measure of seperability. It tells how much model is capable of distinguishing between classes.

Higher the AUC, better the model in predicting 0's as 0's and 1's as 1's. Higher the AUC, better the model in distinguishing between the patients with disease and no disease. The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis. An excellent model has AUC near to the 1 which means it has good measure of separability.

A poor model has AUC near to zero which means that it has the worst measure of separability. When AUC is 0.7, there is 70% chance that model will be able to distinguish between positive class and negative class. All methods used were evaluated based on recall and precision. To compare overall model performance metrics for different classifiers, we collected all the metrics created as shown in Table1 and also the ROC curve for the comparison of models is shown in Fig.3.

Table 1 : Comparing the performance of different algorithms.

Classifier	Accuracy	precision	recall	f1-score	AUC
Logistic Regression	0.62	0.62	0.62	0.57	0.60
Naive Bayes	0.61	0.60	0.61	0.54	0.59
KNN	0.57	0.56	0.58	0.57	0.72
Decision Tree	0.62	0.61	0.62	0.58	0.62
Random Forest	0.62	0.61	0.62	0.60	0.63
AdaBoost	0.62	0.62	0.62	0.57	0.63
Hyperparameter tunung	0.63	0.62	0.63	0.59	0.65

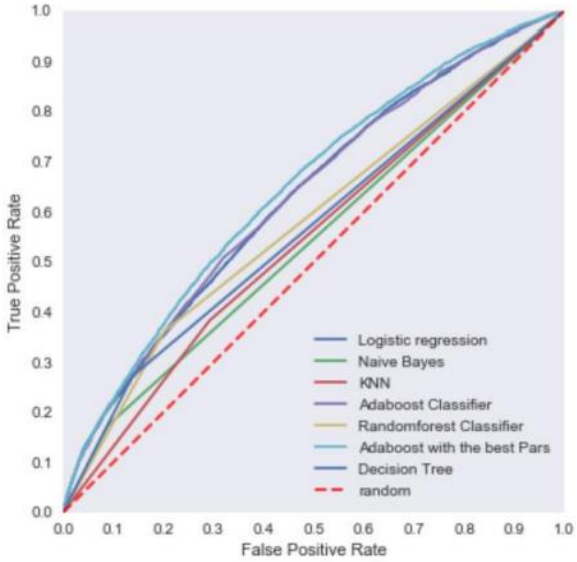

Fig.3: Comparison of models

```
Accuracy of Logistic regression model: 62.3046875
Accuracy of Naive Bayes model: 61.0041920732
Accuracy of KNN model: 57.5552591463
Accuracy of Random forest classification:  61.9998094512
Accuracy of AdaBoosted Classification model:  62.3904344512
Accuracy of Hyperparameter Tuning AdaBoosted Classification model:  62.6810213415
Accuracy of Decision Tree  62.1141387195
```

Fig. 4: Accuracy level of different classifiers

The accuracy levels of AdaBoost after tuning is the best from the result shown above in Fig. 4 about 63%. The accuracy of all other models are similar and ranges between 62-63%.

V. CONCLUSION

Hospital readmission of diabetes patients is an important health care quality measure. Our research suggests that applying a machine learning approach to a larger feature set can improve the prediction. In this project we presented a scheme to identify high risk patients and evaluated different machine learning algorithms. It is found that Adaboost with hyperparameter tuning is optimal for this task. Larger datasets containing medical records of readmitted patients are likely to be helpful for future research. Moreover, this work uncovers the features that are critical in identifying high risk of readmission. Our research targets diabetic patients only. Such analysis

needs to be carried for other top health conditions like Heart diseases, COPD etc .

VI. REFERENCES

[1]. Decision Support in heart disease prediction using naïve bayes-G Subbalakshmi, K Ramesh, MC Rao - Indian Journal of computer Vol. 2 Apr-May 2011.

[2]. Decision Tree Discovery for the Diagnosis of Type II Diabetes Asma A. AlJarullah , 2011 International Conference on Innovations in Information Technology.

[3]. Classification of Heart Disease Using KNN and Genetic Algorithm Volume 10, 2013.

[4]. Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm.M. Akhiljabar, B.L.Deekshatulu, PritiChandra- Volume 10, 2013.

[5]. Design of a Clinical Decision Support Model for Predicting Pneumonia Readmission Jhih-Siou Huang, Yung-Fu Chen , Jiin-Chyr Hsu ,2014 International Symposium on Computer.

Cite this article as :

V R Reji Raj, Mr. Rasheed Ahammed Azad. V, "Predictive Analysis of Heterogenous Data for Hospital Readmission", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 10 Issue 1, pp. 106-112, January-February 2023. Available at doi : https://doi.org/10.32628/IJSRSET231012
Journal URL : https://ijsrset.com/IJSRSET231012