Predicting Early Readmission of Diabetic Patients: Towards Interpretable Models

Mir Moynuddin Ahmed Shibly, Tahmina Akter Tisha, and Md. Mahfuzul Islam Mazumder
Department of Computer Science and Engineering
East West University, Dhaka, Bangladesh
{2016-3-60-057, 2016-3-60-060, 2016-3-60-048}@std.ewubd.com

***Abstract* – Hospital readmission among diabetic patients is a common phenomenon throughout the world. Predicting such patients who have high risk of readmission at the time of discharge even before can help us to provide better healthcare to them. It can minimize the cost associated with readmission too. This study aims at creating a decision support system that can find diabetic patients who are prone to early readmission. To do that, several data mining techniques have been used. Two regular classifiers using the decision tree, and random forest have been developed. After that, two rule-based classifiers using repeated incremental pruning to produce error reduction (RIPPER), and PART algorithms have been developed to provide better interpretability, and understandability of the support system. Between two regular classifiers, the random forest has shown the best performance with 89.5% accuracy and 89.5% recall. And, between rule-based classifiers, PART has demonstrated promising performance with 84.6% accuracy, and 84.6% recall. Using these classifiers, smart and improved healthcare can be ensured.**

Keywords – Rule-based classifiers, early readmission, RIPPER, PART, predictive models, decision support system.

I. INTRODUCTION

Hospital readmission is an important issue in the healthcare system. It can be a matter of inconvenience for patients, doctors, and other stakeholders. Among the patients who have stayed in a hospital for various reasons, diabetic patients are more prone to hospital readmission [1]. A study shows that the readmission of diabetic patients can be managed by better follow-up treatment. To ensure better treatment after discharge, the first step is to identify those who are at risk of getting readmitted early. A statistical decision support system based on data mining techniques that can detect those risky patients early can be helpful to mitigate the risks of mortality. Early predicting of the readmission can also be financially helpful.

Many methods have been used to accomplish the task of early prediction of readmission of diabetic patients. Different researchers have applied different techniques to provide classification based predictive models. Some have used machine learning algorithms [2] like support vector machine [3], random forest [4], neural networks [5], etc. In this domain, rule-based classifiers can be utilized to design an improved predictive support system. As their results are easier to understand than the other 'black-box' systems, rule-based classifiers can also provide better interpretability of the system to the stakeholders. But there is a gap of knowledge of how the rule-based classifiers can perform in early readmission prediction of diabetic patients. This study designs such interpretable rule-based classifiers to predict the early readmission possibility of diabetic patients using Repeated Incremental Pruning to Produce Error Reduction (RIPPER) and PART algorithms. Regular classifiers using decision tree, and random forest algorithms have also been designed.

II. RELATED WORKS

Shaoze Cui et al. [3] have conducted experiments on data about diabetic patient hospital readmission. The aim has been to reduce the readmission rate so that better healthcare can be provided. They have used synthetic minority oversampling technique to reduce the imbalance between the class that has early hospital readmission risk and the class that has less chance of getting readmitted. They have used a hybrid feature selection method combining the filter method, and wrapper method. For the binary classification task, they have proposed an SVM-based method. They have used a genetic algorithm to optimize the parameters of SVM. They have also employed k-fold cross-validation in the training stage. The study has achieved 81.02% accuracy in the testing phase. In another study [4], an intelligent decision support system has been developed using the random forest algorithm and Bayesian network. They have experimented with a series of classification algorithms. They have achieved up to 82.97% accuracy with their works. Additionally, they also have identified the optimal medications for 28 comorbidity combinations to reduce the chance of readmission after discharge. They have also suggested to increase monitoring for the risky patients. Similar studies have also been conducted by Samah Alajmani et al. [6] Bhuvan M S et al. [7], Saumya Salian et al. [8], etc. The different studies have utilized different techniques but to achieve a common goal. All of these studies are aimed to develop predictive systems that can identify the high-risk patients that are more likely to get readmitted to the hospital 30 days after discharge.

III. METHODOLOGY

The objective of this study is to predict the risk of a diabetic patient being readmitted to the hospital again in 30 days after discharge. This work designs a decision support system to predict whether a diabetic patient is needed to be under constant monitoring or not after discharge. In this section of the paper, the methods, and materials used to create such a support system are described.

A. Dataset

In this study, "Diabetes 130-US hospitals for years 1999-2008 Data Set"[9] containing more than 100000 instances from the UCI machine learning repository has been used. They had extracted data from the Health Facts database (Cerner Corporation, Kansas City, MO) based on five criteria. There are 50 features like age, race, weight, the medical specialty of a patient, different diagnoses and medications, etc. in this data set. Each instance of the dataset is labeled with a class having one of the three outcomes – not readmitted, readmitted in less than 30 days, and readmitted after 30 days. Eleven of the features of the data set are numeric, and the rest are nominal.

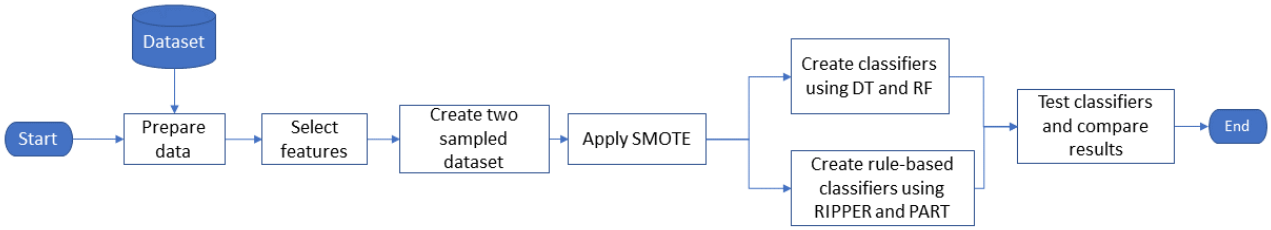

Figure 1: Flow Diagram of the study

B. Data Preparation and Feature Selection

There are few nominal features in the dataset that contain null values. The features that have less than or equal to 2% missing values are replaced with the mode of the corresponding feature. On the other hand, the feature 'medical specialty' has 53% missing values. They are replaced with a new feature value named 'unknown'. Another feature named 'weight' which has 97% missing entries is removed from the dataset. Three other attributes – 'encounter ID', 'patient no', and 'payer code' are also removed. Three categorical variables 'diag_1', 'diag_2', and 'diag_3' contain International Statistical Classification of Diseases and Related Health Problems i.e., ICD9 codes of primary, secondary, and tertiary diagnosis of a patient. These features are mapped to nine major diagnoses as suggested in the original paper of this dataset [9]. In this study, the focus group is the patients with readmission status '<30'. The '>30' status has been merged with the 'NO' group to make it a binary classification problem. In this way, a complete categorical dataset with 45 features with the target feature having binary classes is prepared. The dataset has been needed to be binarized further in some experiments for the compatibility issue of used tools. Additionally, for feature selection, Gini index – a filter model has been used. Gini index calculates the discriminative power of each feature [10]. Using this, 24 features with lower scores are selected for classification.

C. Data Sampling

The initial experiments of this study have been conducted on the complete dataset. But our scope is only diabetic patients. Two sampled datasets have been prepared based on two criteria. With the data points that have a primary diagnosis as diabetes i.e., diag_1 = Diabetes and diabetes_med = Yes, two sampled datasets are created as suggested in these papers [3][4]. In the original dataset and two sampled datasets, there is an imbalance between two target classes. This imbalance in datasets may result in poor performance of the classifiers. And, the classifiers can perform poorly to detect the patients who have a higher chance to get readmitted despite performing well overall. To tackle this imbalance, synthetic minority oversampling technique (SMOTE). In this method, each instance of the minority classes' k nearest neighbors from the same class are calculated. After that, a portion of the neighbors is selected randomly based on the number of oversampling instances needed. Then, for each instance-neighbor pair, a synthetic data point is generated and added to the training set.

D. Classifiers

After preparing data, and feature selection, two classifiers have been developed using decision tree and random forest algorithms. The decision tree algorithm continuously divides the work area by plotting lines into sub-areas until a specific class emerges. Random forest is an ensemble learning method which is widely used in regression and classification problems. A random forest is nothing but a combination of many decision trees. In this method, k decision trees are created using training data to form a forest, and for the test data, the majority voting technique is followed. The term forest came from the idea that to create each decision tree, a random set of attributes is selected. The individual trees are not developed with all features [11]. Random forest resolves the problem of overfitting in decision tree classifiers.

E. Rule-Based Classifiers

As the objective of this study states, in this stage, three rule-based classifiers have been developed. The rule-based classifiers have better interpretability than the typical classifiers. How a regular trained model predicts a class in a real-life environment can be less understandable by the common stakeholders who use the system. But if there are some rules upfront supporting a classifier, then the working mechanism behind a decision can be perceived by the stakeholders. Rules can be generated from a decision tree. But the number of rules generated by a decision tree is enormous. That is why algorithms specially designed for rule-learning are used. Three algorithms have been used to generate rules – apriori algorithm, RIPPER, and PART algorithm. Here, the first one is not a classifier but an associator. And, the other two are classifiers. In this study, a modified rule generation framework has been used for apriori. We are looking for rules that can predict the classes. First, the frequent itemsets that have either of the two classes are filtered. After that, only the rules of types that have either of two classes as the consequence are considered for association rules. If these rules satisfy the minimum confidence, then they are the desired rules mined.

RIPPER is a sequential covering algorithm. It adds three improvements over the Incremental Reduced Error Pruning (IREP) algorithm. IREP starts pruning the created rule right after it was developed. In basic IREP, the training dataset is divided into two sets naming 'growing set' and 'pruning set'. The growing set has two-third of the training data, and the pruning set has the other one-third. For the growing set, using a basic sequential covering algorithm, the best rule is generated. After that, the worth of the rule is calculated on the pruning set based on a metric. Then, a clause from the end of the rule is omitted, and the worth of the reduced rule is computed. If the worth does not decline, the pruning process continues. After a rule is pruned, the training instances covered by the rule are removed, and the next rule generation starts [12]. RIPPER offers few modifications in the IREP procedure. The first improvement is a better metric to

calculate the worth of a rule, the second one is using a different stopping criterion, and the third one is the global optimization step [13]. The last algorithm to create a rule-based classifier is PART [14]. As mentioned earlier, RIPPER globally optimizes the rulesets after rule generation using complex methods. In contrast, PART does not need this global optimization step to generate accurate rules. It uses a typical separate-and-conquer method to build a rule. It follows a sequential covering algorithm. But it differs in how a rule is created. To create each rule, it creates a partial decision tree on current instances. After building the partial tree, one of the leaves that has the highest coverage of training instances is turned into a rule.

IV. Results

This study aims to develop a system that predicts early hospital readmission of diabetic patients. To do that, three versions of the discussed dataset are prepared. After data preparation, decision tree and random forest classifiers are created. Then, three rule-based prediction models have been developed using Apriori, RIPPER, and PART algorithms. To implement the decision tree, and random forest classifiers, scikit-learn library from Python has been used. And, for the rule-based models, Weka – a data mining tool developed at the University of Waikato, New Zealand has been used. For all the classifiers, the datasets have been split into training and testing set having 80% and 20% data respectively. In this section of the paper, the outcomes of the study based on precision, recall, f1-score, and accuracy are presented.

A. Results Analysis

The initial experimentations on the full dataset show that the overall performance of DT and RF is good. The accuracies of classifiers are 88.46% and 89.24% respectively. But we are more interested to detect the patients that have more chance of readmission than the patients have no chance of readmission. Both classifiers have performed poorly to find out the patients that have been readmitted within 30 days after discharge. But after adding synthetic data using SMOTE, the class-wise performance has been improved. Table 1 shows the comparison of the classifiers' class-wise performance. Another aspect is to consider is the impact of feature selection. It has been observed that the classifiers with 45 features, and with 24 features have similar performance. This justifies the employing feature selection strategy before training. A comparison with and without feature selection has been demonstrated in table 2. P, R, and A denote precision, recall, and accuracy respectively. The scope of this study is focused on diabetic patients. To provide a better decision support system for this domain-specific patient, the original dataset has been sampled into two datasets based on primary diagnosis, and diabetes medication. The two sampled datasets contain 8772, and 78363 data points respectively. The experiments on two datasets have yielded a maximum accuracy of 89.5% and 90.6% accuracy respectively.

In the last stage, three rule-based predictive models have been created using Apriori, RIPPER, and PART algorithms. While creating rules from frequent itemsets that are mined using Apriori, minimum support of 0.2 has been used. And 20 rules have been generated with a minimum confidence of 0.7. On the other hand, RIPPER, and PART algorithm have produced 90 and 635 rules respectively on the primary diagnosis-based dataset. Three rules from each predictive model have been presented in table 3 and figure 2 shows the performance comparison between rule-based classifiers and typical classifiers. And, finally, table 4 presents the performance comparison of this study with few existing works which proves that the employed methods of this study can achieve better performance.

TABLE I. Class-wise performance of the classifiers on the full dataset with and without SMOTE

Sampling Technique	*Algorithm*	*Classes*	*Precision*	*Recall*	*Support*
Without sampling	DT	<30	0.30	0.05	2189
		NO	0.90	0.98	18165
	RF	<30	0.48	0.01	2189
		NO	0.89	1.00	18165
SMOTE	DT	<30	0.86	0.90	17930
		NO	0.90	0.85	18234
	RF	<30	0.98	0.91	17930
		NO	0.92	0.99	18234

TABLE II. Impact of feature selection with Gini index

Algorithm	*W/O feature selection*			*With feature selection*		
	P	*R*	*A*	*P*	*R*	*A*
DT	0.884	0.883	**0.883**	0.883	0.882	0.882
RF	0.952	0.950	0.950	0.952	0.950	**0.951**

TABLE III. Few rules generated by rule-based models

Apriori (minconf=0.7)
num_procedures=(-2, 0] number_emergency=(-10, 0] number_inpatient=(-5, 0] → readmitted=NO
discharge_disposition_id=D1 number_inpatient=(-5, 0] time_in_hospital=(0, 5] → readmitted=NO
glipizide=No medical_specialty=unknown metformin=No number_outpatient=(-4, 5] race=Caucasian repaglinide=No A1Cresult=None → readmitted=<30
RIPPER
(medical_specialty = unknown) and (number_inpatient = (0, 5]) and (num_lab_procedures = (40, 60]) and (race = Caucasian) and (A1Cresult = None) and (diag_3 = Circulatory) and (diag_2 = Circulatory) → readmitted=<30
(medical_specialty = unknown) and (num_lab_procedures = (40, 60]) and (race = Caucasian) and (A1Cresult = None) and (diag_3 = Other) and (admission_type_id = Emergency) → readmitted=<30
(number_inpatient = (-5, 0]) and (race = AfricanAmerican) and (num_medications = (0, 10]) and (change = Ch) and (insulin = Steady) → readmitted=NO
PART
number_inpatient = (10, 15] AND medical_specialty = unknown AND A1Cresult = None → readmitted = <30
number_inpatient = (5, 10] AND admission_source_id = AS1 AND race = Caucasian AND medical_specialty = unknown → readmitted = <30
glipizide = Up → readmitted = NO

TABLE IV. Performance comparison with other works

Works	*Test Accuracy*
Shaoze Cui et al. [4]	81.02%
Chinedu I. Ossai et al. [5]	82.97%
The proposed method (RF)	**89.5%**
Proposed method (PART)	**84.6%**

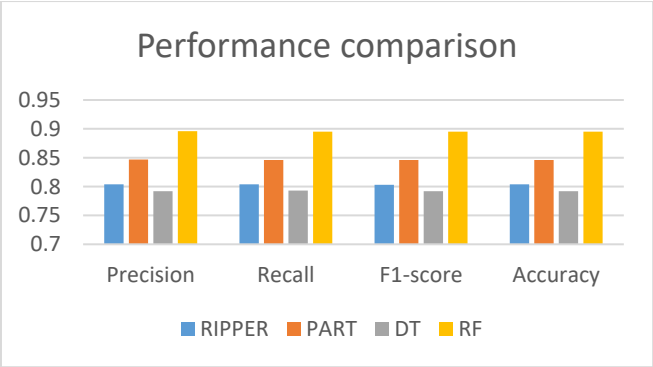

Figure 2. Performance comparison of RIPPER, PART, Decision Tree, and Random Forest

V. Discussion

Predicting early hospital readmission of patients can be helpful to reduce hassles to treat them properly. It can also minimize costs associated with readmission. For diabetic patients, a decision support system that can filter out the patients susceptible to readmission is important too. This study is an effort to develop such a system with better understandability and interpretability. As par the objective of this study, we have created classifiers using various algorithms, and have analyzed their performances. It has been proved that both typical and rule-based classifiers can produce good performing models. Few issues are to be discussed based on the findings from various experiments. One of the major challenges during experimentations was the high imbalance in the dataset we have worked on. To manage this problem, multiple measures have been taken, and the synthetic minority oversampling technique has been proven to work better than other techniques. Generating synthetic data points, and adding them to the original dataset may cause the classifiers to behave differently in the real-world environment. Collecting more data for the minority class can be helpful to develop more sustainable classifiers. The experiments have been carried out to feature selection too. The empirical data shows that, with feature selection, the classifiers can be robust. Among typical classifiers, the random forest algorithm has yielded better performance, and have beaten currently existing works in this domain.

The main research outcome of this study lies in the rule-based predictive models. There have not been many works with those methods in this domain-specific classification tasks. Such models can help the stakeholders in a unique understandable way. The doctors and patients who interact with the system can easily grasp the way the models are deciding by looking at the generated rules. Although, the large number of rules generated by them can be overwhelming for them. Further works on reducing the rules, and exploring ways to simplify them can add an extra dimension to this type of decision support system. Among the rule-based predictive models, the PART algorithm has the best performance. Though it has generated more rules than the RIPPER algorithm, the rules are short and simple. There are some limitations to this study too. To reduce the complexity of the classifiers, the three classes problem has been treated as a binary classification problem. The '>30' class has been merged with the 'NO' class as we are interested in predicting early readmission. Although, this issue could easily be resolved by conducting experiments for three classes. But that would add dis-ambiguity to the interpretation of the developed systems. Another limitation is the lack of using more traditional algorithms. Only the decision tree and random forest algorithm have been applied. The other classification algorithms like Naïve Bayes, SVM, Logistic Regression, etc. could be used for classification purposes, and their comparative analysis could add extra weights to the study. Despite the limitations, the outcome of this study was good, and with the concepts presented by it, hospital readmission of diabetic patients can easily be predicted.

VI. Conclusion

Ensuring better healthcare for risky patients after discharge is important. High-risk diabetic patients should be monitored carefully at the time of discharge. It should be ensured that they do not get readmitted to the hospital soon. Lacking in the follow-up treatment may increase the mortality rate and cost of healthcare too. To prevent these, an automated classifier that predicts the hospital readmission among diabetic patients can be useful. With that view, in this study, the unique idea of rule-based classifiers has been introduced. These types of classifiers can predict the 30 days readmission of a patient in an understandable way. This study has developed a few such systems, and it has been shown with empirical data that they can surely work as a good decision support system.

References

[1] D. J. Rubin, "Correction to: Hospital Readmission of Patients with Diabetes," *Curr. Diab. Rep.*, 2018, doi: 10.1007/s11892-018-0989-1.

[2] M. Alloghani *et al.*, "Implementation of machine learning algorithms to create diabetic patient re-admission profiles," *BMC Med. Inform. Decis. Mak.*, 2019, doi: 10.1186/s12911-019-0990-x.

[3] S. Cui, D. Wang, Y. Wang, P. W. Yu, and Y. Jin, "An improved support vector machine-based diabetic readmission prediction," *Comput. Methods Programs Biomed.*, 2018, doi: 10.1016/j.cmpb.2018.10.012.

[4] C. I. Ossai and N. Wickramasinghe, "Intelligent therapeutic decision support for 30 days readmission of diabetic patients with different comorbidities," *J. Biomed. Inform.*, 2020, doi: 10.1016/j.jbi.2020.103486.

[5] A. Hammoudeh, G. Al-Naymat, I. Ghannam, and N. Obied, "Predicting hospital readmission among diabetics using deep learning," 2018, doi: 10.1016/j.procs.2018.10.138.

[6] S. Alajmani and H. Elazhary, "Hospital Readmission Prediction using Machine Learning Techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 4, 2019, doi: 10.14569/IJACSA.2019.0100425.

[7] M. S. Bhuvan, A. Kumar, A. Zafar, and V. Kishore, "Identifying Diabetic Patients with High Risk of Readmission," Feb. 2016, [Online]. Available: http://arxiv.org/abs/1602.04257.

[8] S. S. D. G. Harisekaran, "Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients," *Int. J. Sci. Res.*, 2015.

[9] B. Strack *et al.*, "Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records," *Biomed Res. Int.*, vol. 2014, 2014, doi: 10.1155/2014/781670.

[10] C. C. Aggarwal, *Data Mining*. Cham: Springer International Publishing, 2015.

[11] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[12] J. Fürnkranz and G. Widmer, "Incremental Reduced Error Pruning," in *Machine Learning Proceedings 1994*, 1994.

[13] W. W. Cohen, "Fast Effective Rule Induction," in *Machine Learning Proceedings 1995*, 1995.

[14] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," *Proc. Fifteenth Int. Conf. Mach. Learn.*, 1998, doi: 1-55860-556-8.