Proceedings of the 2016 Industrial and Systems Engineering Research Conference
H. Yang, Z. Kong, and MD Sarder, eds.

A Meta-heuristic LASSO Model for Diabetic Readmission Prediction

Salih Tutun
Department of Systems Science and Industrial Engineering
Turkish Military Academy, Ankara, Turkey, and Binghamton University, Binghamton, NY

Sina Khanmohammadi, Lu He and Chun-An Chou
Department of Systems Science and Industrial Engineering
Binghamton University, Binghamton, NY

Abstract

Hospital readmission prediction continues to be a highly-encouraged area of investigation mainly because of the readmissions reduction program by the Centers for Medicare and Medicaid services (CMS). The overall goal is to reduce the number of early hospital readmissions by identifying the key risk factors that cause hospital readmissions. This is especially important in Intensive Care Unit (ICU), where patient readmission increases the likelihood of mortality due to the worsening of the patient condition. Traditional approaches use simple logistic regression or other linear classification methods to identify the key features that provide high prediction accuracy. However, these methods are not sufficient since they cannot capture the complex patterns between different features. In this paper, we propose a hybrid Evolutionary Simulating Annealing LASSO Logistic Regression (ESALOR) model to accurately predict the hospital readmission rate and identify the important risk factors. The proposed model combines the evolutionary simulated annealing method with a sparse logistic regression model of Lasso. The ESALOR model was tested on a publicly available diabetes readmission dataset, and the results show that the proposed model provides better results compared to conventional classification methods including Support Vector Machines (SVM), Decision Tree, Naive Bayes, and Logistic Regression.

Keywords
Hospital Readmission, Diabetes, Classification, Metaheuristic Optimization, Regularization

1. Introduction

1.1 Background and Motivations

Nowadays, hospital readmission is one of the leading problems in health-care, mainly because of financial and clinical repercussions. Hospital readmission reduction has become one of the main goals of health-care providers, especially since the CMS introduced the reimbursement penalty for hospital readmissions that occur within 30 days of patient discharge [1]. Hospital readmission is especially problematic for diabetes, since 23% of the annual hospitalizations in the USA are for diabetic patients while they include only 8% of the country's population [2].

In the literature, many researchers have focused on qualitative research methods to explain readmission risk factors [3, 4]. Some studies also assessed different variables for hospital readmission prediction [5]. They mostly used logistic regression because it is easy to calculate the probability of readmission, and to identify the importance of features [6–8]. Moreover, to improve the prediction accuracy, some researchers combined logistic regression with other methods such as artificial neural networks (ANN). However, logistic regression has over-training issue for imbalance data, and combining methods (e.g, ANN, Fuzzy systems) are black-box that cannot provide the probability of readmission and, therefore, are not easy interpretable [9].

In this paper, we propose a hybrid model called Evolutionary Simulating Annealing LASSO Logistic Regression (ESALOR) for hospital readmission prediction. The ESALOR model combines the evolutionary simulated annealing

optimization method with a least absolute shrinkage and selection operator (LASSO) regression approach. The proposed model can be used to analyze the effect of different risk factors on hospital readmission and predict hospital readmission. The proposed model is compared with traditional classification approaches including Support Vector Machines (SVM), Decision Tree (DT), Naive Bayes (NB), and Logistic Regression (LR) to show improvement of models. The organization of the paper is as follows. In Section 2, data preprocessing is explained followed by the details of the proposed hybrid model. In Section 3, results are given to show the performance of proposed model. The paper finishes in Section 4 with a brief conclusion.

2. Materials and Methods

2.1 Data Preprocessing

The diabetes readmission dataset was retrieved from the health facts database, which is a public Electronic Health Record (EHR) data set concerning diabetes patients [10]. The data includes 55 features (such as diagnoses, number of visits, etc.), and the class label is whether or not a certain patient is readmitted within 30 days of discharge. The data set was preprocessed by removing the missing values and applying feature selection methods. We used information from several filters (correlation and information gain), and wrapper (decision tree) feature selection methods to select the most relevant features of the data set. After checking all feature selection methods, 13 features such as discharge disposition, number of inpatients, and diagnosis were selected for our analysis. These selected features will be further filtered in the LASSO component of our proposed hybrid model.

2.2 Artificial Neural Network

Many difficulties, such as the inability to process abnormal data or work with incomplete information, or to solve problems with traditional computer software technologies, can be solved with the Artificial Neural Network (ANN) [11]. The information is contained on the network because information is as precious as the value of connections on the network in ANNs. Users form their own conclusions with the information obtained from samples and after that they are able to make similar decisions on similar cases and process incomplete information on uncertain cases. They are able to make a decision by establishing relevant relationships regarding events after learning them with the help of data. After training the ANN network, it is able to work with incomplete information and give results even if there is incomplete information on recently arrived examples. The information distributed on the network shows that it has a distributed memory. In other words, it is able to work with numeric information [11].

2.3 Support Vector Machine

The Support Vector Machine (SVM) is powerful two category classifier. The algorithm tries to separate hyperplane in the feature space. The algorithm can calculate the distance between every point of independent data looking hyperplane [12]. The minimum one for distances is called margin. The aim of the SVM is to obtain hyperplane of optimum margin, as is seen in Figure 1. In the Figure 1 , you can see the aim of the algorithm with observations on two independent

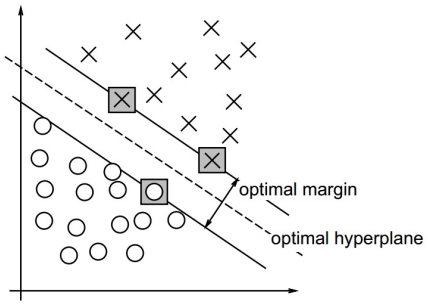

Figure 1: An example of a separable problem in a two dimensional space [12].

variables. However, using linear hyperplane, the algorithm does work well in some cases. Therefore the researchers are using different functions (e.g. radial-based functions, kernel functions). Also, for the misclassification penalty coefficient, the tuning parameters are being used to improve the method in the literature [12, 13].

2.4 Naive Bayes Algorithm

This algorithm is a generative-based model because features are produced independently. It is the simplest model for a machine-learning algorithm. But it also works well for real-world applications. The algorithm considers an unknown target function as $p(y/x)$. In order to learn, $P(y/x)$ is used in training data to calculate $p(x/y)$ and $p(y)$. By using these, we can calculate $p(y/x)$ as you see in Equation (1) [13].

$$P(Y = y_i \mid X = x_k) = \frac{P(X = x_k \mid Y = y_i)p(Y = y_i)}{\sum_j P(X = x_k \mid Y = y_i)p(Y = y_i)} \tag{1}$$

For instance, in order to classify output y, the algorithm is using prior distribution $p(y)$. Afterwards, a sequence of events is made by selecting each event independently from conditional distribution $p(x/y)$. *(An event could be repeated many times)*. Prior distribution $p(y)$ and conditional distribution $p(x/y)$ can be calculated from the training data set. The algorithm can make predictions for the test set by looking at likelihoods from distributions. At the same time, we can estimate parameters by using maximum likelihood or Bayesian estimates. Alternatively, a smoothed estimate can be used [13].

2.5 Logistic Regression

Logistic regression (LR) is approached by learning from function as $p(y/x)$. Y is discrete value, and x is a vector that includes discrete or continuous values. The algorithm is directly estimating parameters from training data.

$$\log \frac{p(x)}{1 - p(x)} = \beta_0 + x\beta \tag{2}$$

$$P(x; b, w) = \frac{e^{\beta_0 + x\beta}}{1 + e^{\beta_0 + x\beta}} \tag{3}$$

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{w_0 + \sum_{i=1}^{n} w_i x_i}} \tag{4}$$

$$P(Y = 0 \mid X) = \frac{e^{w_0 + \sum_{i=1}^{n} w_i x_i}}{1 + e^{w_0 + \sum_{i=1}^{n} w_i x_i}} \tag{5}$$

As you see in Equations (2 - 5), it is like a linear regression model. But the difference is output. For example, in classification, we need to classifies output. Logistic regression classify output by using the above Equations (2 - 5). In this method, there is binary classification as y=1 and y=0. By using a logistic regression equation, the algorithm determines probability. Afterwards, the algorithm classifies the testing value by using threshold. After optimizing the parameters of equations, we can use them to predict output of testing data [14]. The LR is a linear classifier on x value. At the same time, the LR is a function approximation algorithm to use training data to directly estimate $p(y/x)$ [14].

2.6 Evolutionary Simulating Annealing LASSO Logistic Regression (ESALOR) Model

The objective of the proposed model is to optimize coefficients of Logistic Regression (LR) using the evolutionary strategy (ES) and simulated annealing (SA) algorithms and prevent over-training using regularization (Lasso). Simulated annealing is a random-search technique being a trajectory founded using single based optimization. The algorithm searches for the feasible solution space by exploring the neighborhoods of initial solutions [15]. The initial points of the SA algorithm can be identified randomly, however, since searches for nearby points giving initial solutions, it can easily get stuck in local optima. In our framework, we use another meta-heuristic optimization approach named "evolutionary strategy" to identify a good initial solution for simulated annealing. This concept is represented in Figure 2. The randomly initialized SA begins to find solutions from S_0 to S_3, after arriving at S_3, the algorithm tends to accept this point as the optimal solution for decision variables, but it is clearly a local optimum. However, when we initialize the algorithm with solutions found by ES algorithm, the SA algorithm does not get stuck in local optima and can find the optimal solution [16]. By using a hybrid meta-heuristic optimization approach, the coefficients of the model are optimized to find the best model, as is seen in Equation (7).

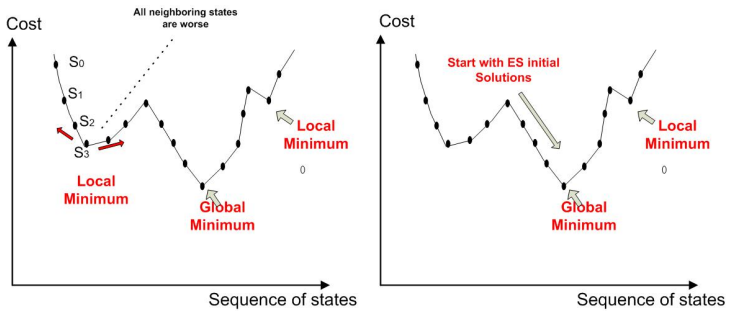

Figure 2: Coupling ES and SA [16]

The typical formulation of logistic regression is shown in Equation (6). This method is used in some of the hospital readmission studies [10, 17], however, the traditional logistic regression model suffers from the over-fitting problem.

Regularization methods have been proven to be an effective approach for solving the overfitting problem by penalizing the absolute of the regression coefficients. The mathematical formulation of LASSO is provided in Equation (7), where N is the number of observations, y_i is the response at observation i, X_i is data point, λ is a non-negative regularization parameter β values are the coefficients of the regression model. This formulation is optimized by using the evolutionary strategy based simulated annealing algorithm because formulation (as an objective function) is not linear with absolute and square values.

$$F_X = \frac{1}{1+e^{-(\beta_{n+1}+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\ldots+\beta_n x_n)}} \tag{6}$$

$$min_{\beta_0,\beta_1,\beta_2,\beta_3,\ldots\beta_n}\left(\frac{1}{2N}\sum_{i=1}^{N}(Y_i-(F_{Xi}))^2+\lambda\sum_{j=1}^{p}\mid\beta_j\mid\right) \tag{7}$$

Considering the provided information, the proposed framework can be summarized in the following steps:

Step 1 Feature Selection: The best subset of features is selected using a combination of filter and wrapper feature selection methods.
Step 2 Formulation: The LASSO-logistic regression formulation of the problem is identified.
Step 3 Initialization: The simulated annealing model is initialized using the evolutionary strategy algorithm.
Step 4 Optimization Level: The parameters (coefficients) of the LASSO model are optimized using a hybrid evolutionary strategy based simulated annealing method. We optimized the parameters of the proposed model.
Step 5 Identifying Solutions: We find the optimal solution by comparing all solutions.
Step 6 Prediction: Hospital readmission of a new patient is predicted using the LASSO model with optimal coefficients.

2.7 Performance Evaluation

The performance of the proposed model is evaluated using four performance criteria including accuracy, recall, precision, and F-measure. Equations (8-11) provide details of this four performance criteria. Among these four performance criteria, the F-measure is generally preferred as it provides a better estimate of the algorithm performance when the testing data set is imbalanced because it compares learning algorithm for each subclass. These measures are based on True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

$$Precision = \frac{TP}{TP+FP} \tag{10}$$

$$F\ Measure = \frac{2TP}{2TP+FP+FN} \tag{11}$$

TP is the number of correct classifications for normal patient detection. FP is the number of incorrect classifications for readmission detection. FN is the number of incorrect classifications for normal patient detection. TN is the number of correct classifications for readmission detection.

3. The Results and Discussion

In this section, initial feature selection and the proposed methods are used to predict hospital readmission rate, and to identify the important risk factors (features). For initial feature selection, correlation-based and information gain-based feature selections are used to select the best subset of features. After checking all feature selection methods, it turns out that the most significant features indicating readmission include discharge disposition, diagnosis and the number of inpatients.

Table 1: Selected features by using Gain ratio based feature selection and Correlation based feature selection

Gain Ratio Feature Selection		**Correlation-based Feature Selection**	
Ranked Rate	**Ranked attributes**	**Ranked Rate**	**Ranked Attributes**
0.0188	Number of Inpatients	0.1059	Number of Inpatients
0.0049	Discharged Disposition ID	0.0786	Discharged Disposition ID
0.0028	Chlorpropamide	0.0587	Patient Number
0.0021	Miglito	0.0513	Time in Hospital
0.0020	Diagnosis 1	0.0303	Encounter ID
0.0012	Diagnosis 3	0.0280	Number of Emergency
0.0017	Diagnosis 2	0.0231	Metformin

For gain ratio based feature selection, as can be seen in Table 1, number of inpatients, discharge disposition, chlorpropamide, miglitol, and diagnosis are very effective for our analysis. For correlation based feature selection, as one can also be seen in Table 1, number of inpatients, discharge disposition, patient number, time in hospital, encounter ID, number of emergencies, and metformin shows significance ranking for readmission. In conclusion, the best initial features, such as number of inpatients, discharge disposition, time in hospital, miglitol, diagnosis, number of emergencies, metrofin and chlorpropamide, are found for the proposed model by looking at feature selection results. The second feature selection is made by using the LASSO shrinkage in the proposed model, as seen in Equation (7). After using the proposed model, other features become zero by penalizing the absolute of the regression coefficients. Therefore, discharge disposition, number of inpatients, diagnosis 1, and diagnosis 2 are selected for training (1/3) level and testing level (2/3) in data.

Table 2: Comparison of ESALOR model with traditional classifiers with testing data.

Methods	**Accuracy**	**Precision**	**Recall**	**F-measure**
SVM	75.11%	0.70	0.75	0.67
ANN	75.85%	0.68	0.75	0.65
LR	74.95%	0.70	0.75	0.65
NB	74.48%	0.68	0.74	0.67
ESALOR	**76.20%**	**0.77**	**0.77**	**0.86**

The results are compared by looking at performance indicators for readmission, and our models are used to make better predictions. Our approach also shows better results than other approaches in the literature comparing four methods. More specifically, for results of the SVM, ANN, LR and NB, as is seen in Table 2, prediction accuracy is founded around 74 % for testing level. Precision and Recall values are less than 0.7 for most methods. At the same time, F-measure values, which need to be more than 0.8, are founded around 0.65 for these methods. Therefore, when using outstanding methods such as the SVM, ANN, LR and NB, prediction performance is inadequate for readmission.

However, our proposed model's performance is much better than other methods such as F-measure. It means that the proposed model works for imbalance data because there is no imbalance learning for each subclass. Therefore, the proposed model performs better in predicting the readmission rate.

4. Conclusion

With the introduction of a reimbursement penalty by the Centers for Medicare and Medicaid (CMS), hospitals have become strongly interested in reducing the readmission rate. In this study, we proposed a hybrid classification framework called Evolutionary Simulating Annealing LASSO Logistic Regression (ESALOR) to improve the classification of readmissions of diabetic patients. The ESALOR model can help health-care providers identify the key risk factors that cause hospital readmission for diabetic patients. By using the identified risk factors, physicians can develop new strategies to reduce readmission rates and costs for the care of individuals with diabetes.

References

1. Centers for Medicare and Medicaid Services, 2016, "Readmissions Reduction Program (HRRP)," retrieved from http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html
2. Centers for Disease Control and Prevention (CDC), and Centers for Disease Control and Prevention (CDC), 2011, "National Diabetes Fact Sheet: National Estimates and General Information on Diabetes and Prediabetes in the United States," retrieved from http://www.familydocs.org/f/CDC
3. Long, T., Genao, I., and Horwitz, L. I., 2013, "Reasons for Readmission in an Underserved High-risk Population: a Qualitative Analysis of a Series of Inpatient Interviews," BMJ open, 3(9),e003212.
4. Strunin, L., Stone, M., and Jack, B., 2007, "Understanding Rehospitalization Risk: Can Hospital Discharge be Modified to Reduce Recurrent Hospitalization?," Journal of Hospital Medicine, 2(5), 297–304.
5. Cooper, G. S., Sirio, C. A., Rotondi, A. J., Shepardson, L. B., and Rosenthal, G. E., 1999, "Are Readmissions to the Intensive Care Unit a Useful Measure of Hospital Performance?," Medical Care, 37(4), 399–408.
6. Garrison, G. M., Mansukhani, M. P., and Bohn, B., 2013, "Predictors of Thirty-day Readmission Among Hospitalized Family Medicine Patients," The Journal of the American Board of Family Medicine, 26(1), 71–77.
7. Hasan, O., Meltzer, D. O., Shaykevich, S. A., Bell, C. M., Kaboli, P.J., Auerbach, A. D., Wetterneck, T. B., Arora, V. M., Zhang, J., and Schnipper, J. L, 2010, "Hospital Readmission in General Medicine Patients: a Prediction Model," Journal of general internal medicine, 25(3), 211–219.
8. van Walraven, C., Dhalla, I. A., Bell, C., Etchells, E., Stiell, I. G., Zarnke, K., Austin, P. C., and Forster, A. J., 2010, "Derivation and Validation of an Index to Predict Early Death or Unplanned Readmission After Discharge from Hospital to the Community," Canadian Medical Association Journal, 182(6), 551–557.
9. Liu, Y., Zayas-Castro, J. L., Fabri, P., and Huang, S., 2014, "Learning High-dimensional Networks with Nonlinear Interactions by a Novel Tree-embedded Graphical Model," Pattern Recognition Letters, 49, 207–213.
10. Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., and Clore, J. N., 2014, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," retrieved from http://dx.doi.org/10.1155/2014/781670
11. Oztemel, E., 2006, "Yapay sinir ağları, (Artificial neural networks)" Papatya Publishing, 2nd Edition, Istanbul, Turkey.
12. Cortes, C., and Vapnik, V., 1995, "Support Vector Networks," Machine Learning, 20(3) 273–297.
13. Dai, W., Brisimi, T. S., Adams, W. G., Mela, T., Saligrama, V., and Paschalidis, I. C., 2015, "Prediction of Hospitalization Due to Heart Diseases by Supervised Learning Methods," International Journal of Medical Informatics, 84(3), 189-197.
14. Mitchell, T. M., 2016, "Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression," retrieved from http://www.cs.cmu.edu/ tom/mlbook/NBayesLogReg.pdf
15. Kirkpatrick, S., 1984, "Optimization by Simulated Annealing: Quantitative Studies," Journal of Statistical Physics, 34(5-6), 975–986.
16. Tutun, S., Chou, C. A., and Canıyılmaz, E., 2015, "A New Forecasting Framework for Volatile Behavior in Net Electricity Consumption: A Case Study in Turkey," Energy, 93, 2406–2422.
17. Walsh, C., and Hripcsak, G., 2014, "The Effects of Data Sources, Cohort Selection, and Outcome Definition on a Predictive Model of Risk of Thirty-day Hospital Readmissions," Journal of Biomedical Informatics, 52, 418-426.