2014 13th International Conference on Machine Learning and Applications

A Comparison of Supervised Machine Learning Techniques for Predicting Short-Term In-Hospital Length of Stay Among Diabetic Patients

April Morton[a]*
NCSR Demokritos
Athens, Greece
april.m.morton@gmail.com

Eman Marzban*
Cairo University
Giza, Egypt
eman.marzban@eng.cu.edu.eg

Georgios Giannoulis, Ayush Patel
Rajender Aparasu and Ioannis A. Kakadiaris
University of Houston
Houston, Texas, USA
{aktorionreg, ayushpatel2008}@gmail.com,
{raparasu, ikakadia}@central.uh.edu

***Abstract*—Diabetes is a life-altering medical condition that affects millions of people and results in many hospitalizations per year. Consequently, predicting the length of stay of in-hospital diabetic patients has become increasingly important for staffing and resource planning. Although statistical methods have been used to predict length of stay in hospitalized patients, many powerful machine learning techniques have not yet been explored. In this paper, we compare and discuss the performance of various supervised machine learning algorithms (i.e., multiple linear regression, support vector machines, multi-task learning, and random forests) for predicting long versus short-term length of stay of hospitalized diabetic patients.**

Keywords*—*Supervised Machine Learning; Support Vector Machines; Support Vector Machines Plus; Random Forests; Multi-Task Learning; Diabetes; In-Hospital Length of Stay Prediction

I. Introduction

Diabetes is a common life-altering autoimmune disease that affects a growing percentage of the population each year [1]. According to the Congressional Diabetes Caucus, a total of 25.8 million people have diabetes and the number of annually hospitalized patients in the US has grown from 2.8 million in 1988 to 5.8 million in 2009 [2]. Moreover, in 2007, the number of hospitalizations totaled 24.3 million days and cost between $1,853 and $2,281 per day, resulting in a cost per day that was more than 2.3 times higher than that for hospitalized patients without diabetes [2].

Due to the growing number of hospitalized diabetic patients, predicting the average length of stay (LOS) has become increasingly important for both resource planning and effective admission scheduling [3]. Obtaining LOS estimates is useful for planning future bed usage, determining specialists for patients with multiple diagnoses, determining health insurance schemes and reimbursement systems in the private sector, planning discharge dates for elderly patients, and allowing families to better plan for the return of their relatives [4].

Although methods have been explored for predicting LOS in hospitalized patients [3], [5], [6], [7], [8], [9], [10], none has assessed and compared the performance of a variety of models across diabetic patients. This paper presents results of an empirical comparison of five supervised learning algorithms applied to diabetic patient records from a large well-known medical database. We evaluate the performance of multiple linear regression (MLR), support vector machines (SVM), support vector machines plus (SVM+), multi-task learning (MTL), and random forests (RF) for predicting long versus short-term length of stay of hospitalized diabetic patients.

II. Related Work

Predicting LOS in hospitalized patients has been studied since the late 1960s [4]. One of the first researchers to tackle this problem, Gustafson [6], compared and evaluated five techniques for prediction of LOS based on a sample of eight inguinal herniotomy patients. Three of the techniques provided users with point estimates based on physicians' subjective opinions, while the other two provided probability distributions over all lengths of stay based on empirical data. Unfortunately, the small sample size reduced the validity of the given models.

Recently, researchers have tackled the LOS prediction problem using statistical and supervised machine learning algorithms. Walczak *et al.* [9] used neural networks to predict the level of illness and length of stay in trauma patients and found that the combination of the backpropagation and fuzzy ARTMAP produced optimal combined results. Kulinskaya *et al.* [8] compared many linear regression-based and maximum likelihood-based models and found that truncated maximum likelihood (TML) had the best fitness value. Liu *et al.* [7] predicted LOS by using a combination of linear and logistic regression models based on a dataset of hospitalizations from 17 hospitals in Northern California. They found that including the Laboratory Acute Physiology Score (LAPS) and Comorbidity Point Score (COPS) greatly improved model performance. Azari *et al.* [10] proposed a multitiered data mining approach for Predicting Hospital Length of Stay (PHLOS) where training sets were created by taking samples close to the cluster centers obtained from k-means clustering. It was found that sampling the data using the k-means clustering method was much more effective than random sampling without clustering.

Most recently, Patel *et al.* [11] compared the performance of several combinations of variables from the Nationwide Inpatient Sample database of 2009 [12] to predict in-hospital mortality and LOS in diabetic patients. Using multiple linear

[a]April Morton is currently affiliated with Oak Ridge National Laboratory.
*Authors made equal contributions.

978-1-4799-7415-3/14 $31.00 © 2014 IEEE
DOI 10.1109/ICMLA.2014.76

regression, they concluded that the best combination of variables for predicting LOS was age, gender, race, insurance, admission type, and the APR-DRG severity measure (described in Section III.B). Unfortunately, even the best performing model exhibited poor performance (an R-squared value of 0.169 for the multiple linear regression model), likely due to the non-linear nature of the problem.

III. Methodology

Though several studies have implemented and assessed models for the prediction of LOS in hospitalized patients, Patel *et al.* [11] have been the only researchers to assess one of these models for diabetic patients. Furthermore, though their assessments have provided an adequate starting point for LOS prediction in diabetic patients, they have only explored one model (multiple linear regression) and have obtained relatively poor results.

In order to more thoroughly explore the potential of predictive methods for estimating LOS in hospitalized diabetic patients, we implemented, compared, and assessed five supervised learning algorithms applied to a subset of diabetic patients from a large well-known medical database. Rather than predicting the exact LOS for hospitalized patients, we predict short-term vs. long-term LOS of each patient, where a less-than-three-day LOS is considered to be short-term. We choose a threshold of less than three days because the distribution of the LOS drastically changes after two days, suggesting that an LOS greater than two days is less common and hence requires special planning. We use the same $p = 6$ features used in Patel's best performing model [11] and evaluate the performance of multiple linear regression (MLR), support vector machines (SVM), support vector machines plus (SVM+), multi-task learning (MTL), and random forests (RF) using the area under the curve (AUC), accuracy (ACC), and f-score (FS) measures. MLR is chosen because it was found to be the best performing method by Patel *et al.* [11], while the remaining four are chosen due to their popularity and suitability for the problem under consideration.

In the following subsections, we describe in greater detail our selected learning algorithms and the dataset.

A. Learning Algorithms

In this section, we introduce each supervised machine learning technique along with a brief description of our choice of algorithms and parameters.

Random Forests (RF): The RF method is used for classification and regression and provides predictions by aggregating results from a large number of decision trees [13]. Several variants of RF have been developed over the years [14], [15] and each depends on the way individual trees are constructed, the procedure used to generate data for the construction of individual trees, and the way predictions of each tree are aggregated to produce final predictions.

In our experiments, we use the original RF method proposed by Breiman *et al.* [14] which constructs each tree from a bootstrap sample drawn with replacement from the original dataset. This method uses the Decrease of Gini Impurity (DGI) as a splitting criterion, selects the splitting predictor from a randomly selected subset of predictors, and aggregates predictions using majority voting. We randomly select $\lceil \sqrt{p} \rceil = \lceil \sqrt{6} \rceil$ predictor variables for each split as suggested by Strobl [16] and aggregate our predictions over 100 decision trees. We observed that these experimental settings minimize the error.

Support Vector Machines (SVM/SVM+): SVM classifiers have been extensively used in various classification problems due to their mathematical foundation and the very good results they produce in various, unrelated fields [17]. The basic intuition behind SVM is to find a hyperplane that discriminates between two classes (implementations for more than two classes are available), but with an extra constraint that the margin between classes should be maximized. The instances that are the closest to the hyperplane are called support vectors. In our experiment, we use SVM with a Gaussian kernel and a parameter $c = 4$.

SVM+ is an extension of SVM and is typically used in conjunction with the Learning Using Privileged Information (LUPI) model developed by Vapnik *et al.* [17]. The goal of the LUPI model is to improve performance by effectively using knowledge that is available for training but not testing. In our experiment we use the SVM+ algorithm with a Gaussian kernel and parameters $c = 1$ and $g = 0.25$. All unused features in the dataset are used as privileged information.

Multi-Task learning (MTL): The MTL method is a variant of another related machine learning method called Single-Task Learning (STL). STL aims to break large problems into small, reasonably independent sub-problems that are learned separately and then recombined. Though STL often produces acceptable results, Curuana [18] argues that sometimes the methodology is counterproductive because it ignores the information contained in the training samples of other tasks drawn from the same domain. Consequently, several algorithms for MTL [19], [20], [21] were developed to improve the performance of learning algorithms by learning classifiers for multiple tasks jointly.

In our experiment, we use the Lasso MTL method [21] to estimate LOS in diabetic hospitalized patients. The Lasso method imposes a sparsity condition on the regression coefficients for the predictors (the features) and thus provides a means for both indicating feature strength in the prediction process and accelerating computation time. We divide the tasks based on hospital regions (Northeast, Midwest, South, West) and empirically chose an l_1 regularization parameter of $\lambda = 0.01$.

Multiple Linear Regression (MLR): The goal of MLR is to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to the observed data. Linear regression models are often fit using the least squares approach, but may also be fit by minimizing the "lack of fit" in some other norm, as with least absolute deviations regression, or by minimizing a penalized version of the least squares loss function, as in ridge regression and kernel ridge regression [22].

In our experiment, we determine the coefficients using the

least-squares model, which finds the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. In order to transform our real-valued predictions to binary values we convert our prediction to 1 if the prediction is less than 3 and 0 otherwise.

B. Dataset

In our experiment, we use the HCUP Nationwide Inpatient Sample database from 2009 [12]. It is the largest all-payer hospital discharge database containing approximately 8 million patient records. Each record contains several features including demographics (age, race, gender), hospital information (location, number of beds, teaching vs. non-teaching), admission type, number of diagnoses (up to 25), health insurance status, total hospital charges, risk/severity measures, and length of stay. In our experiment, we use the same six features used in the best performing model found by Patel [11] *et al.* These include Age Category (under 18 years of age, between 18 and 65 years of age, over 65 years of age), Indicator of Sex (Male or Female), Race (White, Black, Hispanic, Asian/Pacific Islander, Native American, Other), Expected Primary Payer (Medicare, Medicaid, Private Insurance, Uninsured, Other), Admission Type (Emergency, Urgent, Elective, Newborn, Delivery, Trauma Center, Other), and the All Patient Refined DRG (APR-DRG) measure [23]. The sixth feature, the APR-DRG measure, is assigned using software developed by 3M Health Information Systems that considers factors such as the severity of the illness and the risk of mortality of the patient [24].

IV. Experimental Results

We created a smaller version of the dataset by strategically selecting 10,000 patient records from the HCUP Nationwide Inpatient Sample Database described in Section III.B. To construct the dataset we first chose the same number of patients from different regions of the USA (Northeast, Midwest, South, West) to keep our model general. Also, since diabetes is a disease that occurs in both sexes we choose to keep an equal number of males and females. Moreover, we have observed that in the entire database the number of patients with short and long stay is roughly equal in size so we choose to keep the classes of equal size. In addition, we choose the 25 largest hospitals (according to number of diabetic discharges), with 100 patients from each hospital. We use 5-fold cross-validation to obtain five trials for testing where, during each trial, we use 8,000 cases to train each model and 2,000 cases to test each model. For each method, we compute the mean and standard deviation of the area under the ROC curve (AUC), accuracy (ACC), and f-score (FS). A one-way ANOVA test, followed by Bonferroni's post-hoc comparisons test ($\alpha = 0.05$), is performed on each set of metrics to determine if the differences among the results for each model are significant.

Table 1 and Figure 1 depict scores for each algorithm on each of the three metrics. In addition, Figure 2 depicts the ROC curves for each model for the trial with highest AUC. Each of the highest scoring metrics in Table 1 is highlighted in bold. All differences in mean ACC are significant except for those between SVM vs. SVM+ and vs. RF. The mean FS of MTL is significantly different than the mean FS of MLR and SVM+, but the mean FS of RF and SVM are not significantly different than the FS of any of the other models. All mean AUC are significantly different except for the differences between the mean AUC of SVM+ and SVM.

Given the significance of the results, we can conclude that SVM+ achieves both the highest mean ACC and highest mean AUC, followed by RF, MTL and MLR. This particular ranking of algorithms is likely due to the nature of each of the learning algorithms combined with the problem under consideration. More specifically, MLR likely performs poorly due to the non-linear nature of the problem. MTL achieves slight improvement because it improves generalization by learning tasks in parallel using a shared representation. However, this improvement is minimal because linearity is assumed. A more significant improvement is observed for the RF algorithm because the method is non-parametric and generally has more flexibility when splitting the data. The SVM+ algorithm outperform the previous three algorithms because a non-linear Gaussian kernel is used and additional information that is available during training but not during testing is considered.

In addition, it is worth noting that though MLR produces the lowest mean ACC, it achieves the highest mean FS. This is likely due to the fact that the MLR resulted in zero false negatives, causing the FS to increase despite a low mean ACC.

TABLE 1. Performance Results

Method	ACC	FS	AUC
RF	0.65 ± 0.01	0.64 ± 0.01	0.70 ± 0.01
SVM	0.66 ± 0.03	0.63 ± 0.10	0.74 ± 0.03
SVM+	**0.68 ± 0.01**	0.65 ± 0.03	**0.76 ± 0.01**
MTL	0.55 ± 0.01	0.55 ± 0.02	0.56 ± 0.01
MLR	0.50 ± 0.00	**0.66± 0.00**	0.45 ± 0.01

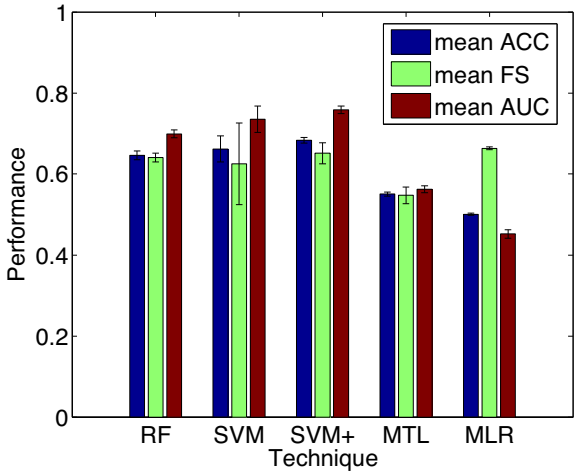

Fig. 1. Standard error bars and mean accuracy (ACC), f-score (FS), and area under the ROC curve (AUC) for the RF, SVM, SVM+, MTL, and MLR supervised machine learning algorithms.

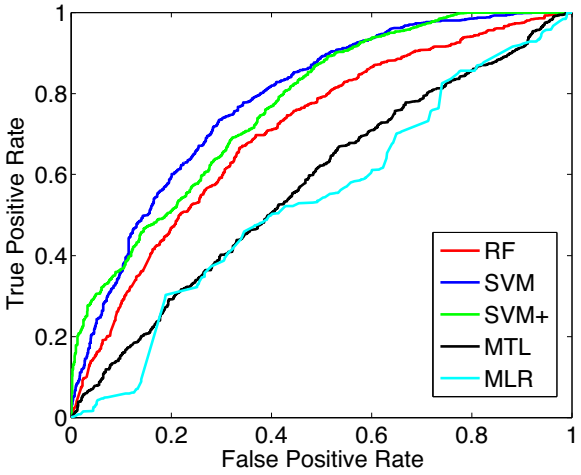

Fig. 2. ROC Curve corresponding to the set of results with maximum AUC for the RF, SVM, SVM+, MTL, and MLR supervised machine learning algorithms.

V. Conclusion and Future Work

In this paper, five supervised learning algorithms were applied to a subset of diabetic patient records from a large well known medical database. Using the Age Category, Indicator of Sex, Race, Expected Primary Payer, Admission Type, and APR-DRG variables we predicted short-term vs. long-term LOS for each patient, where short-term is defined as less than 3 days. We evaluated the performance of the MLR, SVM, SVM+, MTL, and RF supervised machine learning techniques using the AUC, ACC, and FS measures.

Overall, the results indicated that the SVM+ method is most promising for predicting short-term LOS in hospitalized diabetic patients, followed closely by the RF technique. This is likely due to the use of a non-linear Gaussian kernel as well as privileged information that is not available in the training of the other models.

Future work includes conducting a more thorough feature selection method, implementing the MTL learning methodology with other algorithms, increasing the size of the training and testing datasets, and investigating other machine learning algorithms such as artificial neural networks, logistic regression, naive bayes, decision trees, and bagged trees.

Acknowledgments

The authors graciously acknowledge the funding and support of the International Research Centered Summer School (IRSS).

References

[1] R. J. Cornall, J.-B. Prins, J. A. Todd, A. Pressey, N. H. DeLarato, L. S. Wicker, and L. B. Peterson, "Type 1 diabetes in mice is linked to the interleukin-1 receptor and lsh/lty/bcg genes on chromosome 1," *Nature*, vol. 353, no. 6341, pp. 262–265, 1991.

[2] Centers for Disease Control and Prevention (CDC), "National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the united states, 2011," *Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention*, vol. 201, 2011.

[3] G. H. Robinson, L. E. Davis, and R. P. Leifer, "Prediction of hospital length of stay," *Health Services Research*, vol. 1, no. 3, p. 287, 1966.

[4] V. Panchami and N. Radhika, "A novel approach for predicting the length of hospital stay with dbscan and supervised classification algorithms," in *Proceedings of the Fifth International Conference on the Applications of Digital Information and Web Technologies*. IEEE, 2014, pp. 207–212.

[5] P. R. Hachesu, M. Ahmadi, S. Alizadeh, and F. Sadoughi, "Use of data mining techniques to determine and predict length of stay of cardiac patients," *Healthcare Informatics Research*, vol. 19, no. 2, pp. 121–129, 2013.

[6] D. H. Gustafson, "Length of stay: prediction and explanation," *Health Services Research*, vol. 3, no. 1, p. 12, 1968.

[7] V. Liu, P. Kipnis, M. K. Gould, and G. J. Escobar, "Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables," *Medical Care*, vol. 48, no. 8, pp. 739–744, 2010.

[8] E. Kulinskaya, D. Kornbrot, and H. Gao, "Length of stay as a performance indicator: robust statistical methodology," *IMA Journal of Management Mathematics*, vol. 16, no. 4, pp. 369–381, 2005.

[9] S. Walczak, W. E. Pofahl, R. J. Scorpio *et al.*, "Predicting hospital length of stay with neural networks." in *FLAIRS Conference*, 1998, pp. 333–337.

[10] A. Azari, V. P. Janeja, and A. Mohseni, "Predicting hospital length of stay (phlos): A multi-tiered data mining approach," in *Proceedings of the 12th International Conference on Data Mining Workshops*. IEEE, 2012, pp. 17–24.

[11] A. Patel, M. Johnson, and R. Aparasu, "Predicting in-hospital mortality and hospital length of stay in diabetic patients," *Value in Health*, vol. 16, no. 3, pp. A17–A17, 2013.

[12] Agency for Healthcare Research and Quality: Healthcare Cost and Utilization Project (HCUP). (2006-2009) HCUP Nationwide Inpatient Sample (NIS).

[13] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 6, pp. 493–507, 2012.

[14] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. CRC press, 1984.

[15] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: A conditional inference framework," *Journal of Computational and Graphical statistics*, vol. 15, no. 3, pp. 651–674, 2006.

[16] C. Strobl, J. Malley, and G. Tutz, "An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests." *Psychological methods*, vol. 14, no. 4, p. 323, 2009.

[17] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Networks*, vol. 22, no. 5, pp. 544–557, 2009.

[18] R. Caruana, *Multitask Learning*. Springer, 1998.

[19] J. Baxter, "A model of inductive bias learning," *J. Artif. Intell. Res.*, vol. 12, pp. 149–198, 2000.

[20] S. Ben-David, J. Gehrke, and R. Schuller, "A theoretical framework for learning from a pool of disparate data sources," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2002, pp. 443–449.

[21] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[22] A. O. Sykes, "An introduction to regression analysis," 1993.

[23] Agency for Healthcare Research and Quality. (2014) NIS description of data elements. [Online]. Available: http://www.hcup-us.ahrq.gov/db/nation/nis/nisdde.jsp

[24] R. F. Averill, N. Goldfield, B. Steinbeck, T. Grant, J. Muldoon, A. Brough *et al.*, "All patient refined diagnosis related groups (APR-DRGs)," *Version 20.0*, vol. 15, pp. 98–054, 2003.