Evaluation of Classification Methods for the Prediction of Hospital Length of Stay using Medicare Claims data

Dimitrios Zikos  
Heracleia lab, CSE, UT Arlington,  
Box 19015, 500 UTA Blvd.,  
zikos@uta.edu

Konstantinos Tsiakas  
Heracleia lab, CSE, UT Arlington,  
Box 19015, 500 UTA Blvd.,  
konstantinos.tsiakas@mavs.uta.edu

Fadiah Qudah  
CSE, UT Arlington  
Box 19015, 500 UTA Blvd.,  
fadiah.qudah@mavs.uta.edu

Vassilis Athitsos  
Associate Professor  
CSE, UT Arlington  
Box 19015, 500 UTA Blvd.,  
athitsos@uta.edu

Fillia Makedon  
Professor  
CSE, UT Arlington  
Box 19015, 500 UTA Blvd.,  
makedon@uta.edu

ABSTRACT
In this paper, we investigate the performance of a series of classification methods for the prediction of the hospital Length of Stay (LOS), based on two temporally sequential clinical scenarios. We used a 2012 Medicare Provider Analysis and Review (MedPar) dataset, which contains records of Medicare beneficiaries who used inpatient hospital services. Our subset included 300,000 randomly selected cases. During the prepossessing we added new features and linked our data with external datasets, using common key identifiers. In the first scenario our goal was to predict the LOS using a subset of information which is readily available to the clinician upon the patient admission, while the second scenario assumes that there is available additional data (information on the patient diagnosis and clinical procedures). For our experiments we used three different classifiers: Naïve Bayes, AdaBoost and C4.5 Decision tree, for two different LOS cut-off points (4 day and 12 day hospital stay). The overall performance of our classifiers was ranging from fair to very good. On the other hand the true positive rate, that is the correct classification of the long hospital stays, was low, with an exception of Naïve Bayes, which demonstrated significantly better performance in the second scenario. Our results indicate that Naïve Bayes may be used for the prediction of the in-hospital LOS. Our analysis also indicates that the MedPar data combined with other data resources has the potential to provide a good basis for robust prediction analytics in hospitals.

Categories and Subject Descriptors

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org or Publications Dept., ACM, Inc., fax +1 (212) 869-0481.PETRA14, May 27-30, 2014, Island of Rhodes, Greece.
Copyright 2014...ACM 978-1-4503-2746-6/14/05S$15.00

General Terms

Keywords
Length of Stay, Healthcare, Machine learning, Naïve Bayes, Decision Trees, Prediction

1. INTRODUCTION
Length of stay is an important indicator of the quality of care and patient safety and while it varies among cases with different conditions and complications, is also dependent, to a degree, upon priorities of the healthcare system and hospital strategies. LOS is very often investigated by researchers since there are many implications related with both the quality of care and the cost of care. LOS is driven by specific hospital policies, and this often happens due to the need to reduce costs.

There are many studies exploring the determinants of hospital LOS. Husted et al. [1] found that the among hip and knee replacement cases, various factors like the age, sex, marital status, co-morbidities, time between surgery and mobilization, and others, influence a series of outcomes, including the LOS. In a multivariate analysis among pediatric liver transplant recipients [2], an increased hospital stay was found to be associated with factors such as infants less than one year of age, fulminant liver failure, government insurance, and transplant era. In another example, among patients hospitalized with decompensated congestive heart failure [3], LOS was partially related to patient demographics, severity of illness, management modalities, response to treatment, and administrative data.

Reduced stay times are related to better health outcomes and improved quality of care as well. Mowery et al [4] found that during the initial evaluation in Emergency Departments (ED), the hospital mortality rate increased for each additional hour any patient spent in the ED. Quick response to an emergency is equally important and a possible variability in the LOS may also be related with the responsiveness of a hospital. In a large-scale multivariate study [5], researchers found that patients who underwent craniotomy or drainage of hematoma soon after arrival
had half the likelihood of death compared with cases where the surgery happened more than 4 hours after ED arrival.

All this evidence indicates the importance of the LOS for the high standard healthcare provision. Health professionals would want to know about the expected length of stay for their patients, and in addition they would like to be provided with recommendations about targeted interventions which would help achieve an appropriate hospital LOS. They would finally like to know about the LOS for similar patient cases, which has been related with better outcomes. In this paper we want to explore how accurately it is possible to predict whether a hospital stay would be shorter or longer than specific cut-off points, using enormous datasets of Medicare patients. Towards this direction, we investigated the performance of a series of classification methods aiming to predict the hospital LOS and to compare the performance of different algorithms, based on two different scenarios:

- The first scenario assumes that the patient has just been admitted to the hospital, therefore the available information about the patient is limited.
- The second scenario assumes that the patient has already been hospitalized and the clinicians already know about the diagnosis, the clinical procedures and other in-hospital acquired information.

The algorithms which are discussed in this paper can be potentially integrated into the Electronic Medical Records. The data we have been using is either acquired during the patient admission (scenario 1) and the hospital care (scenario 2) or can easily become available through other data resources. We removed data which are available prospectively in the real hospital world (after the discharge of the patient), because we wanted our proposed algorithms to be potentially used concurrently with the use of the Electronic Medical Records during the clinical care. This is, for example, the reason we included no data of financial nature, since these are usually prospectively calculated after the patient is discharged.

Since we use a comprehensive dataset with Medicare cases coming across the whole US, a secondary goal of our study is to explore the potential of the use of similar datasets in order to provide tools for health professionals for the predictions and decision making in order to ensure a high standard of health services and a patient safety driven hospitalization.

2. METHODOLOGY

We used well known data mining techniques to determine the LOS of hospitalized patients. Data mining analyses large data sets to find relationships and summarize the information contained in a meaningful and useful manner [6]. Data mining can be divided in unsupervised learning and supervised learning [7-9]. The former, clusters data by determining their similarity, helping patterns to emerge. The latter, defines classification rules in the form of models, which are then used to classify new, unknown data. Supervised learning includes classification, regression, time series analysis and prediction [10] and was used in our experiments. For the data preprocessing we used the statistical analysis software SPSS [11] and for the classification experiments we used Weka 3.7 [12].

We preprocessed the dataset and created deriving attributes using SPSS syntax editor. For example, we added new features describing the in-hospital acquired diseases, by using the MedPar POA (Diagnosis was Present on Admission) attributes. We also created dummy features from categorical variables where each category was representing different aspects of care.

The data we used for the data analysis is a 2012 Medicare Provider Analysis and Review (MEDPAR) file, which contains records for the 100 percent of Medicare beneficiaries who used hospital inpatient services during the aforementioned year. The records are stripped of most data elements that will permit identification of beneficiaries. The hospital is identified by the six-position Medicare bill number. We randomly selected 300,000 cases for our experiments. The features which were included can be classified into three categories with temporal sequence in terms of the data acquisition and availability. As shown in Fig.1, t₀ represents the patient admission time slot and this is when the initial information becomes available. T₁, T₂ represents the patient hospitalization period. T₃ represents the time when the patient was discharged.

Two scenarios have been conceptualized, each one assuming the availability of different feature sets. Since the first scenario assumes that the patient has just been admitted to the hospital, the information that is known about the patient is the following:

(i) Patient demographics (age, gender, ethnicity)
(ii) Information about their hospital (National Provider ID)
(iii) Admission information (type of admission, admission day/season, source of admission).

The second scenario assumes that the patient has already been hospitalized for a few days and we already know the diagnosis, procedures and other in-hospital acquired information. The second scenario is temporally consequent to the first one.

For both scenarios, we used two different LOS cut-off points. The first cut-off point is equal to 4 days and the second cut-off point is 12 days. We want to predict the LOS being more or less than 4 days, since a LOS equal to 4 days falls approximately into the 50% percentile of the LOS distribution. Regarding the second cut-off point, a LOS equal to 12 days falls approximately into the 90% percentile of the LOS distribution, therefore instances above this cut-off point represent a hospitalization longer than the 90% of our cases. For all our experiments we compared the performance of three classifiers: Naïve Bayes, AdaBoost and C4.5 (J48 in Weka) decision tree classifier. We used the Weka environment for Knowledge Analysis, version 3.7. Fig. 2 presents our experimental design.

Naïve Bayes is a probabilistic classifier based on Bayes’ theorem. It is called “naïve” because it inherently assumes the conditional independence between different features. Its main advantages are its simplicity due to the need for only one probability multiplication and these results in low computational complexity. Additionally, it requires a relatively small training set since it only calculates the feature outcome frequency pairs. These characteristics make this classifier capable of handling large data sets very quickly and attain accurate results. The inherent independency assumption of the classifier is often left unsatisfied with real data; however the effect on the classification accuracy may be minimal [13].
We also used AdaBoost, a well-known classifier, for our experiments [14]. AdaBoost uses the boosting technique, combining a number of weak classifiers with the weighted majority voting scheme. Classifiers with higher accuracy during the training phase acquire higher weights in the final classifier. AdaBoost has been shown to outperform most classifiers (including SVMs) in the literature when applied in the medical domain. Such examples include its application on epiluminescence microscopy (ELM) [15] and MRI images [16].

Finally, we also used the C4.5 decision tree classifier, which builds a tree structure in a top-down, divide-and-conquer manner [17]. Data is divided into classes, and the most informative features become nodes of the tree. The construction of the tree using “if-then” criteria constitutes the training phase. Its main advantage is the visualization of the decision making procedure, but the visualization is not useful when the feature set is big. Pruning methods may be applied in such cases, to discard the least important nodes and to make the tree easier to interpret and to reduce over-fitting.

3. RESULTS

For both scenarios we performed experiments with the goal to predict the LOS using Naïve Bayes classifier, AdaBoost, and the J48, which is the Weka implementation of the C4.5 decision tree algorithm. Among all 300,000 cases included in our analysis, the average LOS was found to be 5.79 days (SD=9.4) and the median was equal to 4 days.

3.1 Scenario 1: ‘The patient has just been admitted’

3.1.1 Class: Length of Stay > 4 days

In our first scenario, we assume that health professionals only have the admission information available. Using ‘LOS more than 4 days’ as our class, Naïve Bayes classified the 66.4% of our instances correctly. More specifically, the classifier performed well in the case of short (less than 4 day) hospital stays, at a rate of 83.8%. On the contrary, Naïve Bayes misclassified the 58.4% of the instances with a LOS>4 days.

We performed the same experiment with the use of AdaBoost meta-classifier, this time. With our class being ‘LOS more than 4 days’ the overall classification performance was found to be 60.5%. Specifically, while AdaBoost classified almost all instances with LOS<4 days correctly (99.8%), it failed to maintain an acceptable performance for the instances with a LOS>4 days.

Finally, the C4.5 decision tree algorithm, classified the 65.3% of our instances correctly. Specifically, the 87.5% of the short (LOS<4 days) have been correctly classified, but most of the instances with a LOS >4 days were misclassified.

3.1.2 Class: Length of Stay > 12 days

We ran experiments with a same setup as of above, only changing our class, which in this case is the ‘LOS more than 12 days’. With this experiment, we wanted to investigate how the admission information alone can predict whether a patient would stay in the hospital for an extended period, longer than that of the 90% of our instances. Naïve Bayes classified the 91.2% of our instances correctly. More specifically, the classifier performed really well in the case of hospital stays with LOS<12 days, at a rate of 96.7%. On the contrary, the classifier showed a weak classification performance for the prediction of long hospital stays.

AdaBoost classified the 91.2% of our instances correctly. While the algorithm classified virtually all instances with LOS<12 days correctly, again, it performed very poorly, during the classification of the instances with a LOS>12 days, which is what clinicians need to know. Finally, the C4.5 decision tree algorithm classified the 91.6% of our instances correctly. In that case, while the algorithm classified the 99.1% of the instances with LOS<12 correctly, it failed to do so with the very long stays.
3.2 Scenario 2: ‘The clinical information is known’

3.2.1 Class: Length of Stay > 4 days

Our second scenario is based on the assumption that we know the clinical information for the patient, on top of the already known admission information. The patient may have been hospitalized for few days, which is in many cases a sufficient period for the diagnoses to be established and for some of the clinical procedures and interventions to take place.

Table 1. Overall performance and precision of the three classifiers

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall performance (%)</td>
<td>% of correctly classified instances=1</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>LOS&gt;4 days</td>
<td>66.4%</td>
</tr>
<tr>
<td></td>
<td>LOS&gt;12 days</td>
<td>91.2%</td>
</tr>
<tr>
<td>Ada Boost</td>
<td>LOS&gt;4 days</td>
<td>60.5%</td>
</tr>
<tr>
<td></td>
<td>LOS&gt;12 days</td>
<td>91.1%</td>
</tr>
<tr>
<td>J48 (Weka Implementation of C4.5)</td>
<td>LOS&gt;4 days</td>
<td>65.3%</td>
</tr>
<tr>
<td></td>
<td>LOS&gt;12 days</td>
<td>91.6%</td>
</tr>
</tbody>
</table>

Using ‘LOS more than 4 days’ as our class, Naïve Bayes classified the 74.1% of our instances correctly. More specifically, the classifier performed well in the case of short (less than 4 day) hospital stays, at a rate of 82.9%. This time, the Naïve Bayes performed better during the classification of the instances with LOS>4 days, in comparison with the first scenario. Specifically a percentage of 61.5% of these instances were classified correctly, which is a big improvement in comparison.

Similarly to the first scenario, we performed the same experiment with the use of AdaBoost meta-classifier. With our class being ‘LOS more than 4 days’ the overall classification performance was found to be 68.3%. Specifically, AdaBoost classified the 85.9% of our instances with LOS<4 days correctly, and although it did not maintain an acceptable performance for the instances with a LOS>4 days, the performance was somehow improved compared with the first scenario, but still not close enough to a desired performance level.

Finally, the C4.5 decision tree algorithm classified the 72% of our instances correctly, showcasing a better overall performance. More specifically, the algorithm classified the 86% of our instances with LOS<4 days correctly and the 51.8% of the longer than 4 day hospital stays. The positive predictive value of the algorithm lies in between Naïve Bayes and AdaBoost.

3.2.2 Class: Length of Stay > 12 days

We ran the two experiments changing our class to ‘LOS more than 12 days’, in order to predict very long hospital stays.

In this experiment, Naïve Bayes classified the 87.8% of our instances correctly. More specifically, the classifier performed really well in the case of hospital stays with LOS<12 days, at a rate of 90.3%. In this scenario, the classifier performed better in comparison to the first scenario, since it classified correctly the 61.6% of instances with a LOS>12 days.

AdaBoost, classified the 91.1% of our instances correctly. As with the first scenario, the algorithm classified virtually all but 12 instances with LOS<12 days correctly, but again it demonstrated a very poor performance, during the classification of the instances with a LOS>12 days.

AdaBoost, classified the 91.1% of our instances correctly. As with the first scenario, the algorithm classified virtually all but 12 instances with LOS<12 days correctly, but again it demonstrated a very poor performance, during the classification of the instances with a LOS>12 days.

Finally, in the case of C4.5, the overall classification accuracy was found to be 92%. Although the shorter than 12-day hospital stays were classified correctly at a rate of 98.3%, the instances with a LOS>12 days were not classified correctly for the majority of our
instances. Table 1 and Fig. 3 & 4 summarize the results of our experiments.

Figure 3. Overall performance of the algorithms for Scenarios 1 and 2

Figure 4. Positive predictive value of algorithms for Scenarios 1 and 2

4. CONCLUSIONS

The results of our experiments indicate that the inclusion of data about the diagnoses and the clinical procedures (Scenario 2) improves the classification performance of Naïve Bayes, whether we want to predict stays longer than the 50% of our instances (LOS>4 days), or very long hospital stays, spanning from 12 days and above. In any case, although the algorithm classifies the instances with LOS<4 and LOS<12 with a very high success ratio -and this is the reason for the high overall performance- it fails though to provide acceptable performance for the prediction of instances above the cut-off points. We need to pinpoint, though, that the addition of the clinical features into our model (scenario 2) has led to an increase of the classification accuracy by 20%, giving a positive predictive value of 60% and even higher. We believe that the availability of more clinical features in our dataset, like the prescribed medication, re-admission ratio and meta-data about the health providers (hospital size, bed capacity etc.), would have helped the classifier demonstrate a better performance.

Adaboost, on the other hand failed to correctly classify instances with LOS>12, in both scenarios. Only in the case of the 4 day cut-off point though, in our second scenario, the classifier classifies almost half of our instances with a class value=1 correctly. Finally, J48 demonstrated the highest overall performance in most of our experiments due to the high specificity ratio. The classifier though did not come close to the positive predictive value of Naïve Bayes. Both classifiers, showcase high sensitivity, since they classify instances with LOS lower than our cut-off points successfully, in both scenarios. Especially in the case of Adaboost, the sensitivity is in the majority of our experiments between 98-100%.

Focusing on the data mining techniques, Naïve-Bayes classifier has many applications of different areas and is used to represent the probabilistic knowledge and use it for classification. In [18], a Naïve-Bayes classifier has also been used to predict length of stay based on a variety of data such as demographics, lab tests, complaint reports etc., reaching an accuracy of 85.8% of correctly classified instances, while K-Nearest Neighbor underperforms with an accuracy rate of 62.6%. Other algorithms, such as the MLP (Multilayer back propagation) algorithm, have been shown to demonstrate an even higher accuracy.

Predicting the length of stay of impatient cases based on various types of information and data has been a challenging task with different approaches. Some researches propose a method that analyzes the temporal trends of LOS, extracting information about the healthcare quality and the hospital activity [19]. Other approaches lie on the field of data mining techniques and use algorithms such as Naïve Bayes, K-Nearest Neighbors and others [18]. In some cases, researchers combine statistical and data mining methods with physicians’ medical experience and opinions.

The results of our experiments indicate that the Length of Stay can potentially be predicted, provided that the Medicare claims data are combined with more detailed clinical data, which are readily available during the clinical process. It is possible that the MedPar data combined with other data resources, will provide a good basis for prediction analytics in hospitals. An estimation of the LOS would facilitate the healthcare provision, but primarily, we believe that the hospital administration would benefit, since the management of hospital resources and the planning of the discharge procedure would become more informed.

In addition, any information about the estimation of the expected LOS can be used as an indirect evaluation of the hospital revenue, since each case is being compensated based on the Diagnosis Related Codes (DRGs) which a-priory integrates an average LOS based on national norms.

Predicting the length of stay in hospitals is obviously influenced by the nature of the data that are used. There are many approaches to the selection of the features to be used for the formation of the model. In a recent study [18], four different levels of clinical information were used, such as demographics, admission and clinical data, including symptomatic data before and after a surgical operation, in order to predict and measure the impact of demographics on the length of stay. Moreover, other approaches combine similar data with a selection of specific criteria as in [20], where researchers linked changes in serum creatinine with the total cost and the length of stay.

MedPar data are easily accessible and readily available for analysis purposes. In our study we wanted to demonstrate the potential and limitations of the use of such data to predict in-hospital LOS. We strongly believe, that given these features are combined with other data resources with more detailed clinical information and additional hospital attributes, the integration of prediction capabilities into the ongoing clinical process will
become a reality. Such integration has many implications regarding (i) the patient evaluation and clinical assessment and (ii) the achievement of higher compliance to regulations. Such functionalities may eventually become an integral part of future EMR applications.

5. ACKNOWLEDGMENTS

a. This material is based upon work supported by the National Science Foundation under Grants No. NSF-CNS 1035913, NSF-CNS 0923494.

b. We would like to thank the ‘Gordian Health Management Group’ for providing to us the 2012 MedPar dataset for the purpose of this study.

6. REFERENCES


