Healthcare Data Mining: Predicting Hospital Length of Stay (PHLOS)

Ali Azari, Information System department, University of Maryland, Baltimore County (UMBC), Baltimore, MD, USA

Vandana P. Janeja, Information System department, University of Maryland, Baltimore County (UMBC), Baltimore, MD, USA

Alex Mohseni, Emergency Medicine Associates, Potomac, MD, USA

ABSTRACT

A model to predict the Length of Stay (LOS) for hospitalized patients can be an effective tool for measuring the consumption of hospital resources. Such a model will enable early interventions to prevent complications and prolonged LOS and also enable more efficient utilization of manpower and facilities in hospitals. In this paper, the authors propose an approach for Predicting Hospital Length of Stay (PHLOS) using a multi-tiered data mining approach. In their approach, the authors form training sets, using groups of similar claims identified by k-means clustering and perform classification using ten different classifiers. The authors provide a combined measure of performance to statistically evaluate and rank the classifiers for different levels of clustering. They consistently found that using clustering as a precursor to form the training set gives better prediction results as compared to non-clustering based training sets. The authors have also found the accuracies to be consistently higher than some reported in the current literature for predicting individual patient LOS. Binning the LOS to three groups of short, medium and long stays, their method identifies patients who need aggressive or moderate early interventions to prevent prolonged stays. The classification techniques used in this study are interpretable, enabling them to examine the details of the classification rules learned from the data. As a result, this study provides insight into the underlying factors that influence hospital length of stay. They also examine the authors’ prediction results for three randomly selected conditions with domain expert insights.

Keywords: Classification, Clustering, Data Mining, Healthcare, Length of Stay, Predictive Models

INTRODUCTION

A model that help to predict a patient’s Length of Stay (LOS) during a single visit, the time from hospital admission until discharge, can be an effective tool in hands of health care providers to (a) plan for preventive interventions, (b) to improve health services, and (c) to manage the hospital resources more efficiently. The productivity of hospitals drop significantly in two situations: First, if the hospital is in short supply for required resources such as manpower and facilities. Second, if the hospital is over equipped and the supply is more than the demand. Both of these situations occur due to significant fluctuations in hospital occupancy, which seriously restricts the efficient scheduling for resource allocation and management. With

DOI: 10.4018/jkdb.2012070103
an accurate estimation of how long patients will stay, the hospital can plan for a better bed management and more efficient resource utilization (Gustafson, 2002). Predicting the possible discharge dates can lead to better estimation of available bed hours, which finally results to higher average occupancy and less waste of hospital resources (Robinson, Davis, & Leifer, 1966). On the other hand, hospitals are continuously being expected to do more with ever diminishing resources. As Medicare reimbursement trends towards ‘pay for performance’, tying payments to efficiency, hospitals stand to lose a lot of money if they cannot predict and prevent excessively long LOS. Therefore, predicting the patients who need the most aggressive early intervention, and those who require a moderate amount of intervention to prevent prolonged LOS seems to be critical. Just as hospitals have created rapid response teams of clinicians to treat patients with decompensated disease, we believe hospitals could create rapid response care teams to intervene on any patient predicted to have a prolonged LOS. It remains to be seen, from a domain perspective, how much of an impact early focused intervention can have on preventing complications and prolonged LOS.

To build our models we use classification techniques for following reasons: 1) unlike regression, the classification techniques do not assume correlation among attributes in the dataset, and 2) in the current literature, the results of predictions made by classification techniques shows higher accuracy than those produced by regression techniques. In this paper we present a Multi-tiered Data mining approach for Predicting Hospital Length of Stay (PHLOS), to reduce the uncertainty associated with the length of stay for hospitalized patients. Specifically, we make the following contributions:

- We identify groups of similar hospital claims using clustering, where the number of clusters is determined based on the disease conditions identified in the literature (Escobar et al., 2008) or by using the Charlson index (Charlson et al., 1987), which provides the general categories of the diseases. We utilize these groups to create our training sets that resulted in predictions with high accuracy. Our accuracy exceeds that of several other models for predicting LOS in the current literature (Gustafson, 2002; Woods et al., 2000; Clark & Ryan, 1968; Abbi et al., 2007; Liu et al., 2006);
- Creating training sets based on k-means clustering, we remove the noisy data from the training sets and used a stratified sampling technique to avoid overfitting the classifiers by selecting representative training sets;
- We provide a combined measure of performance to statistically evaluate and rank the performance of classifiers for different levels of clustering, we reported the statistical significance of the ranking and comparisons performed in this study;
- Through clearly defined LOS classes, we provide a method to predict which patients need the most aggressive early interventions, and which patients require a moderate amount of interventions to prevent complications and long stays;
- We validate our findings with a domain expert in the area of Emergency Medicine, Dr. Mohseni, one of the coauthor’s of this paper. We examine our prediction results for three randomly selected conditions, namely Heart4, Renal2 and pregnancy with domain expert insights.

In our approach, we first create clustering (using K-means clustering) of preprocessed data. We use different numbers of K, for example, corresponding to the conditions or corresponding to Charlson index (Charlson et al., 1987). Then we use these cluster assignments to identify the training data for the classifier. Additionally, we extract the test set, which is non-overlapping from the training sets. We run different classifiers such as Bayesnet, SVM, JRIP, J48, Bagging to name a few, to predict LOS. We extract several performance measures for the classifiers such as Accuracy, Kappa