Discovering the hidden risk factors: An empirical evaluation of incorporating feature-learning methods into a risk model framework using the OMOP CDM

Jenna M. Reps, PhD, Patrick B. Ryan, PhD
Janssen Research and Development, Raritan, NJ

Abstract

Longitudinal observational data can be used to develop risk models by learning condition occurrence patterns that commonly precede an illness. These patterns may consist of latent features that are not directly observed, but rather, they are inferred. In this research, two common approaches for feature learning are investigated, namely deep learning/boosting and matrix factorization. The empirical performances of different risk model frameworks, incorporating various feature learning methods, are evaluated across a range of outcomes on the OPTUM CDM database. It was found that deep learning slightly outperformed lasso logistic regression, indicating the benefits of learning feature interactions, but none of the risk models performed satisfactorily,

Introduction

The rise of digital infrastructure has resulted in large quantities of medical longitudinal observational data becoming available. Furthermore, these datasets are rapidly expanding. This presents the opportunity to learn new medical information from the data; however, methods that can overcome issues with the data are required to fully utilise it. For example, the data could be used to develop high performance personalised risk models that could accurately identify patients who are at high risk of developing certain illnesses. Such models would enable preventative initiatives to be implemented to lower the patients' risks and improve healthcare. However there are often technical issues limiting the analysis of big data[1]. The main issues include the size and sparseness of the data.

To develop high performance risk models it is likely that manual feature engineering or data-driven techniques for feature learning are required. At present, risk models generally rely on expert specified features (basic feature engineering). Using only a small number of expert specified features ignores a large amount of data that could be used to improve the model performance. For example, by considering a patient's complete medical history it may be possible to learn predictive temporal patterns that would otherwise be ignored by conventional risk models. These patterns could be learned by implementing data-driven feature learning techniques such as matrix factorization[2] or deep learning[3]. Aside from often improving model performance, feature learning has the added advantage of being adaptable to different illnesses, so the same methodology could be widely applied to learn predictive features for any illness without requiring experts to specify the features.

Deep learning is a supervised learning technique that learns to infer classes based on learning feature representations. It has been successfully applied to image recognition where it is capable of identifying parts of an object, such as edges or shapes, at different hierarchal levels. Matrix factorization is an unsupervised learning method often used to reduce the dimension of sparsely represented data by identifying latent features. Matrix factorization has been used to improve the performance of risk models[2] and can readily overcome missing data issues. One advantage of matrix factorization is that it can readily incorporate temporal aspect of the data.

The aims of this research are:

1) To empirically evaluate the implementation of deep learning and matrix factorization within a risk model framework using OPTUM CDM data.
2) Provide R code of the risk model framework.

Methods

The risk model framework methodology is presented in Figure 1 below. To extract the cohort data an index date is chosen and all patients within a user specified age range at index are selected if they have either a complete 2-year observational period after the index date or a recording of the outcome during the 2-year interval after index. Patients with less than six years of observation prior to the index date or with a prior recording out the outcome are excluded. For each patient a medical history feature matrix is constructed, with each row representing consecutive 6-month time intervals prior to the index date, each column corresponds to a condition and the entry is 1 if the patient has the condition recorded during the time interval and 0 otherwise. Each patient's outcome label is 1 if they have the outcome recorded within the 2-years follow-up and 0 otherwise. The set of patient feature matrices paired

with their outcomes are used to train a deep neural network for the deep learning framework. For the matrix factorization we implement non-negative matrix factorization to reduce the dimensionality of the data. The matrix factorization is used to predict the risk by feeding the patients' latent features, discovered via matrix factorization, into a classifier (neural network or gradient boosting machine). The code for the risk framework is available online (see poster).

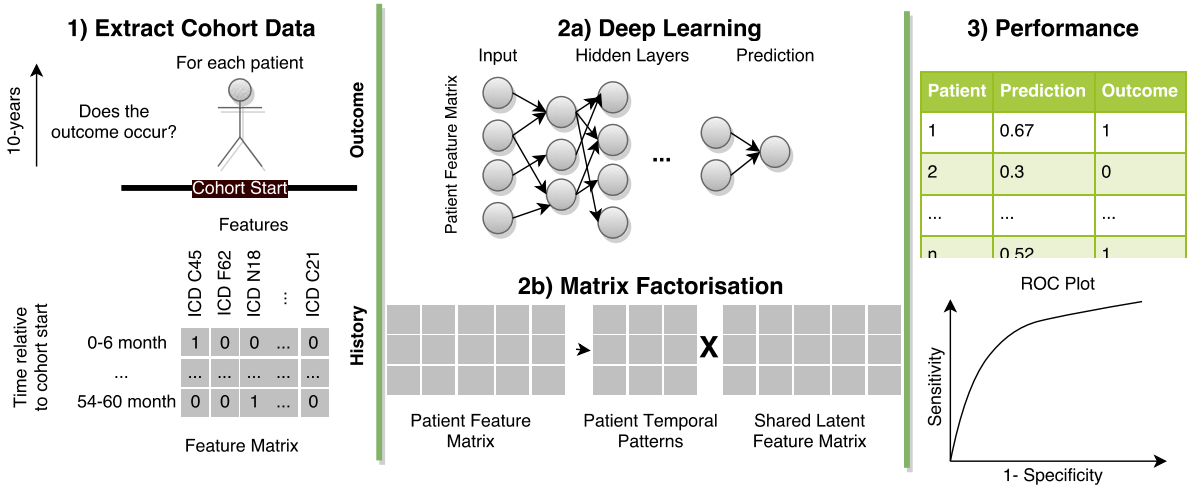

Figure 1. The risk model framework incorporating feature learning via deep learning or matrix factorization.

To compare the frameworks the area under the receiver operating characteristic curves (AUCs) when predicting the risk of OMOP congestive heart failure, OMOP depression and gastro-intestinal hemorrhage are calculated. The complete data is randomly partitioned into a test set (10%) and training set (90%). The hyper-parameters for each risk model are selected based on the results from 5-fold cross validation on the training set.

Results & Discussion

Table 1. The performance of risk model frameworks incorporating the different feature learning methods

Method	*Heart Failure*	*Depression*	*Gastro-intestinal*	*Mean*
Deep Learning Classifier	0.745	0.640	0.641	0.675
Matrix Factorization				
Lasso Logistic Regression	0.734	0.623	0.626	0.661

Table 1 presents the AUC results of each risk model framework applied on the various outcomes. Overall none of the models obtained high AUCs (>0.8). However, as these models only used the 5 year medical condition history and did not include feature such as age or gender, the performance is reasonable. The lasso logistic regression generally performed just as well deep learning and was faster to train. One reason for deep learning only slightly outperforming the other models may be due to the complexity of training deep learning models often resulting in a suboptimal model. The performance shows deep learning created more suitable latent features than matrix factorization. One possible explanation is that the matrix factorization learns latent feature independently of the outcome, whereas the deep learning latent features adapt to the outcome, giving deep learning an advantage. Non-negative matrix factorization seemed to perform poorly, however this may not be true for other matrix factorization methods that impute and learn temporal patterns[2], and this may be worth investigating in the future.

Conclusion

In this paper we investigated the risk model performance when incorporating supervised (deep learning) and unsupervised (matrix factorization) feature learning into the framework. Deep learning performed slightly better than regularized logistic regression across the three outcomes. This shows that feature learning may be a possible solution to personalizing risk models, but more advanced modeling such as implementing ensembles may be required to obtain the desired AUCs. Possible areas of future work include implementing a convolutional neural network approach or including prescriptions and demographic features into the feature learning framework.

References

1. Verleysen M, François D. The curse of dimensionality in data mining and time series prediction. Computational Intelligence and Bioinspired Systems 2005;758-770.
2. Zhou J, et al. From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining 2014;135-144.
3. Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks 2015;61:85-117.