Comprehensible Explanation of Predictive Models

D

Marko Robnik-Šikonja
University of Ljubljana, Slovenia

INTRODUCTION

In many areas where machine learning methods are applied the practitioners and users of produced prediction models are interested in comprehensible explanation of their predictions. Unfortunately, the best performing predictive models do not offer an intrinsic introspection into their decision processes or provide explanations of their prediction. This is true for Support Vector Machines (SVM), Artificial Neural Networks (ANN), and all ensemble methods (for example, boosting, random forests, bagging, stacking and multiple adaptive regression splines). Approaches that do offer an intrinsic introspection such as decision trees or decision rules do not perform so well or are not applicable in many cases (Meyer et al., 2003). The areas where models' transparency is of crucial importance are for example most of business and marketing applications where the executives are just as interested in the comprehension of the decision process, explanation of the existing and new customers' needs and expectations in a given business case, as in the classification accuracy of the prediction model. The same is true for many areas of business intelligence, finance, marketing, insurance, medicine, science, policy making, and strategic planning where knowledge discovery dominates prediction accuracy.

To alleviate this problem two types of solutions have been proposed. The first type is based on internal working of each particular learning algorithm and exploits its learning process to gain insight into the presumptions, biases and reasoning leading to final decisions. A well-known example of such an approach are random forests for which several visualizations exist mostly exploiting the fact that during bootstrap sampling some of the instances are not selected for learning and can serve as an internal validation set. With the help of this set important features can be identified and similarity between objects can be measured. The second type of explanation approaches are general and can be applied to any predictive model. Examples of this approach are methods EXPLAIN (Robnik-Šikonja & Kononenko, 2008) and IME (Štrumbelj et al., 2009). These two methods are based on efficient implementation of input perturbations. They can explain models' decision process for each individual predicted instance as well as the model as a whole. As both methods are efficient, offer comprehensible explanations, can be visualized, and are readily available in R package ExplainPrediction (Robnik-Šikonja, 2015) they are the focus of this article. Other general explanation methods are discussed in the background Section.

The objective of the article is to explain how EXPLAIN and IME explanation methods work and to show their practical utility in several real world scenarios. The first aim is achieved through explanation of their working principle and graphical explanation of models' decisions on a well-known data set. Two types of explanations are demonstrated, predictions of new unlabeled cases and the functioning of the model as a whole. This allows inspection, comparison, and visualization of otherwise opaque models. The practical utility of the methodology is demonstrated with short description of several applications: in medicine (Štrumbelj et al. 2010), macro economy (Pregeljc et al., 2012) and business consultancy (Bohanec et al., 2015).

DOI: 10.4018/978-1-5225-2255-3.ch181

BACKGROUND

In a typical data science problem setting, users are concerned with both prediction accuracy and the interpretability of the prediction model. Complex models have potentially higher accuracy but are more difficult to interpret. This can be alleviated either by sacrificing some prediction accuracy for a more transparent model or by using an explanation method that improves the interpretability of the model. Explaining predictions is straightforward for symbolic models such as decision trees, decision rules, and inductive logic programming, where the models give an overall transparent knowledge in a symbolic form. Therefore, to obtain the explanations of predictions, one simply has to read the rules in the corresponding model. Whether such an explanation is comprehensive in the case of large trees and rule sets is questionable.

For non-symbolic models there are no such straightforward explanations. A lot of effort has been invested into increasing the interpretability of complex models. A taxonomy of explanation methods and a review of neural network explanation approaches is given by Jacobsson (2005). For Support Vector Machines an interesting approaches is proposed by Hamel (2006). Many approaches exploit the essential property of additive classifiers to provide more comprehensible explanations and visualizations, e.g., (Jakulin et al., 2005) and (Poulin et al. 2006).

Visualization of decision boundaries is an important aspect of model transparency. Barbosa et al. (2016) present a technique to visualize how the kernel embeds data into a high-dimensional feature space. With their Kelp method they visualize how kernel choice affects neighborhood structure and SVM decision boundaries. Schultz et al. (2015) propose a general framework for visualization of classifiers via dimensionality reduction. Goldstein et al. (2015) propose another useful visualization tool for classifiers that can produce individual conditional expectation plots, graphing the functional relationship between the predicted response and the feature for individual instance.

Some explanations methods (including the ones presented here) are general in a sense that they can be used with any type of classification model (Lemaire et al. 2008; Robnik-Šikonja and Kononenko 2008; Štrumbelj et al., 2010). This enables their application with almost any prediction model and allows users to analyze and compare outputs of different analytical techniques. Lemaire et al., (2008) applied their method to a customer relationship management system in telecommunications industry. The method which successfully deals with high- dimensional text data is presented in (Martens and Provost, 2011). Its idea is based on general explanation methods presented here and offers explanation in the form of a set of words which would change the predicted class of a given document. Bosnić et al. (2014) adapt the general explanation methodology to data stream scenario and show the evolution of attribute contributions through time. This is used to explain the concept drift in their incremental model.

Many explanation methods are related to statistical sensitivity analysis and uncertainty analysis (Saltelli et al., 2000). In that methodology sensitivity of models is analysed with respect to models' input. A related approach, called inverse classification (Aggarwal et al. 2010) tries to determine the minimum required change to a data point in order to reclassify it as a member of a different class. A SVM model based approach is proposed by (Barbella et al., 2009).

Another sensitivity analysis-based approach explains contributions of individual features to a particular classification by observing (partial) derivatives of the classifiers prediction function at the point of interest (Baehrens et al. 2010). A limitation of this approach is that the classification function has to be first-order differentiable. For classifiers not satisfying this criterion (for example, decision trees) the original classifier is first fitted with a Parzen window-based classifier that mimics the original one and then the explanation method is applied to this fitted classifier. The method was shown to be practically useful with kernel based classification method to predict molecular features (Hansen et al., 2011).