Explaining Predictive Model Decisions

Marko Robnik-Šikonja
Faculty of Computer and Information Science, University of Ljubljana, Slovenia

Erik Štrumbelj
Faculty of Computer and Information Science, University of Ljubljana, Slovenia

INTRODUCTION

Research in statistics, data mining, pattern recognition, and machine learning is mostly focused on prediction accuracy. As a result, we have many excellent prediction methods. Some of the most successful approaches are Support Vector Machines (SVM), Artificial Neural Networks (ANN), and ensemble methods (for example, boosting and random forests). Regrettably, these approaches do not offer an intrinsic introspection into their decision processes or provide explanations of their predictions. Approaches that do offer an intrinsic introspection such as decision trees do not perform so well or are not applicable in many cases (Meyer et al., 2003). In many areas where machine learning and data mining models are applied, their transparency is of crucial importance. For example, in many business and marketing applications 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.

Recently several general explanation methods have been introduced (Robnik-Šikonja & Kononenko, 2008; Lemaire et al., 2008; Štrumbelj et al., 2009; Baehrens et al., 2010) that are relatively independent of the prediction model, and can be used with all classification models that output probabilities. Here we describe two representatives of them, sharing common idea and background, namely the methods EXPLAIN (Robnik-Šikonja & Kononenko, 2008) and IME (Štrumbelj et al., 2009). We discuss other general methods in the background Section.

The EXPLAIN and IME can explain any prediction model, either transparent, for example, decision trees and rules, or a black box, for example, SVM, ANN, and classifier ensembles. These explanation methods decompose the model's predictions into individual contributions of each attribute. Generated explanations closely follow the learned model and enable its visualization separately for each prediction case and also for the modeled problem as a whole.

We explain how these two explanation methods work and graphically explain models' decisions for new unlabeled cases and the workings of the model as a whole. We demonstrate how this allows inspection, comparison, and visualization of otherwise opaque models. We support this description with two applications, a medical (Štrumbelj et al., 2010) and economical (Pregeljc et al., 2012).

BACKGROUND

In a typical 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

DOI: 10.4018/978-1-4666-5202-6.ch085

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 such as ANN (d'Avila Garcez et al., 2001; Palade et al., 2001). For a good review of neural network explanation methods we refer the reader to Jacobsson (2005). For Support Vector Machines interesting approaches are proposed by Hamel (2006) and Poulet (2004). Many approaches exploit the essential property of additive classifiers to provide more comprehensible explanations and visualizations (Jakulin et al., 2005; Mozina et al., 2004; Poulin et al., 2006). Some explanations methods (including the ones presented here) are more general in a sense that they can be used with any type of classification model (Lemaire et al., 2008; Robnik-Šikonja & Kononenko, 2008; Štrumbelj et al., 2010). This enables their application with almost any prediction model and allows users to analyse and compare outputs of different analytical techniques.

In the context of feature subset selection, attributes are evaluated in (Lemaire et al., 2004) as the difference between the correct and perturbed output, which is similar to EXPLAIN approach to the model level explanation (Robnik-Šikonja & Kononenko, 2008). In (Lemaire et al., 2008) this approach was extended to instance level explanations and was applied to a customer relationship management system in telecommunications industry.

In the context of explaining data-driven classifications of text documents, the main issue is computational efficiency. The method which successfully deals with high- dimensional text data is presented in (Martens & 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.

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 which is what we call model level explanation. Presented visualization of averaged explanations can therefore be viewed as a form of sensitivity analysis. A related approach, called inverse classification (Mannino & Koushik, 2000; 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 notable issue 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).

MAIN FOCUS

General explanation methods can be applied to any classification model which makes them a useful tool both for interpreting models (and their predictions) and comparing different types of models. Such methods cannot exploit any model-specific properties and are limited to perturbing the inputs of the model and observing changes in the model's output (Lemaire et al., 2008, Robnik-Šikonja & Kononenko, 2008; Štrumbelj et al., 2010).

Related Content

Cyber Physical Systems: A Review

Siddhartha Khaitanand James D. McCalley (2014). *Encyclopedia of Business Analytics and Optimization (pp. 574-579).*

www.irma-international.org/chapter/cyber-physical-systems/107260

Automation of System Infrastructure Upgrade or Downgrade Using AI

Chiranjeevi Nageswarababu Jonnalagadda (2023). *Handbook of Research on AI and Knowledge Engineering for Real-Time Business Intelligence (pp. 243-253).*

www.irma-international.org/chapter/automation-of-system-infrastructure-upgrade-or-downgrade-using-ai/321498

Application of Triplet Notation and Dynamic Programming to Single-Line, Multi-Product Dairy Production Scheduling

Virginia M. Mioriand Brian Segulin (2010). *International Journal of Business Intelligence Research (pp. 9-20).*

www.irma-international.org/article/application-triplet-notation-dynamic-programming/43678

The Business Transformation Enterprise Architecture Framework

Antoine Trad (2020). *Handbook of Research on IT Applications for Strategic Competitive Advantage and Decision Making (pp. 309-344).*

www.irma-international.org/chapter/the-business-transformation-enterprise-architecture-framework/262483

Exploring Insurance and Natural Disaster Tweets Using Text Analytics

Tylor Huizinga, Anteneh Ayanso, Miranda Smoorand Ted Wronski (2017). *International Journal of Business Analytics (pp. 1-17).*

www.irma-international.org/article/exploring-insurance-and-natural-disaster-tweets-using-text-analytics/169217