Influence of Domain and Model Properties on the Reliability Estimates’ Performance

Zoran Bosnić, University of Ljubljana, Slovenia
Igor Kononenko, University of Ljubljana, Slovenia

ABSTRACT

In machine learning, the reliability estimates for individual predictions provide more information about individual prediction error than the average accuracy of predictive model (e.g. relative mean squared error). Such reliability estimates may represent decisive information in the risk-sensitive applications of machine learning (e.g. medicine, engineering, and business), where they enable the users to distinguish between more and less reliable predictions. In the authors’ previous work they proposed eight reliability estimates for individual examples in regression and evaluated their performance. The results showed that the performance of each estimate strongly varies depending on the domain and regression model properties. In this paper they empirically analyze the dependence of reliability estimates’ performance on the data set and model properties. They present the results which show that the reliability estimates perform better when used with more accurate regression models, in domains with greater number of examples and in domains with less noisy data.

Keywords: Accuracy, Prediction Error, Regression, Reliability, Reliability Estimate

INTRODUCTION

When using supervised learning for modeling data, we aim to achieve the best possible prediction accuracy for the unseen examples which were not included in the learning process (Kononenko & Kukar, 2007). For evaluation of the prediction accuracies, the averaged accuracy measures (e.g. the relative mean squared error). However, such measures provide no information about the expected error of an individual prediction for an unseen example.

Examples of areas, where such reliability estimates for the individual predictions present an important benefit (Crowder, Kimber, Smith, & Sweeting, 1991), are the risk-sensitive areas, where acting upon predictions may have financial or medical consequences (e.g. medical diagnosis, stock market, navigation, control applications). For example, in the medical diagnosis, the use of individual prediction reliability estimates allows the physicians to decide more easily whether to trust the automatically predicted patient’s diagnosis or not.

In our previous work we proposed and compared eight reliability estimates (based on...
five approaches) for the model-independent reliability estimation for individual examples in regression (Bosnič & Kononenko, 2008b). We evaluated performance of the proposed reliability estimates using 5 regression models (regression trees, linear regression, neural networks, support vector machines and locally weighted regression) and on 15 testing domains. The empirical evaluation of our approach showed that our approaches provide good estimation of the prediction error for individual examples. As such, they offer a tool which improves the safety of implementing the automatically generated decisions in the risk-sensitive areas.

In this article we extend the previous evaluation with the sixth regression model (random forests) and larger number of testing datasets (27). Additionally, we perform experiments using artificial data sets and empirically analyze the dependence of reliability estimates’ performance on the data set and model properties (noise in data, number of examples, model accuracy).

The article is organized as follows. First, we introduce reliability estimates that were proposed and empirically evaluated in our previous work. After describing their evaluation protocol and presenting their performance on regression models and 27 testing domains, we empirically analyze the dependence of their performance on the domain and regression model properties. We focus on analyzing impacts of three factors: the accuracy of the regression model, the number of examples in the training set, and the training data noise level. We finish by providing conclusions and ideas for the further work.

**RELATED WORK**

In order to enable users of classification and regression models to gain more insight into the reliability of individual predictions, various methods aiming at this task were developed in the past. Some of these methods were focused on extending formalizations of the existing predictive models, enabling them to make predictions with their adjoined reliability estimates. The other group of methods focused on the development of model-independent approaches, which are more general, but harder to analytically evaluate with individual models. In the following, we present the related work from the both groups of approaches.

The idea of reliability estimation for individual predictions originated in statistics, where confidence values and intervals are used to express the reliability of estimates. In machine learning, the statistical properties of predictive models were utilized to extend the predictions with adjoined reliability estimates, e.g. with support vector machines (Gammerman, Vovk, & Vapnik, 1998; Saunders, Gammerman, & Vovk, 1999), ridge regression (Nouretdinov, Melluish, & Vovk, 2001), and multilayer perceptron (Weigend & Nix, 1994). Since these approaches are bound to a particular model formalism, their reliability estimates can be probabilistically interpretable, thus being the confidence measures $0$ represents the confidence of the most inaccurate prediction and $1$ the confidence of the most accurate one). However, since not all approaches offer probabilistic interpretation, we use more general term, the reliability estimate, to name the measure that provides information about the trust in accuracy of the individual prediction.

In contrast to the previous group of methods, the second group is more general and model-independent (not bound to particular model). These methods utilize the model-independent approaches to find their reliability estimates, e.g. local modeling of prediction error based on input space properties and local learning (Birattari, Bontempi, & Bersini, 1998; Giacinto & Roli, 2001), using the variance of bagged models (Heskes, 1997), or by meta-predicting the leave-one-out error of a single example (Tsuda, Rätsch, Mika, & Müller, 2001). Inspired by transductive reasoning (Vapnik, 1995), Ku kar and Kononenko (2002) proposed a method for estimation of the classification reliability. Their work introduced a set of reliability measures which successfully separate correct and incorrect classifications and are independent of the learning algorithm. We later adapted this
Decision Rule Extraction for Regularized Multiple Criteria Linear Programming Model
DongHong Sun, Li Liu, Peng Zhang, Xingquan Zhu and Yong Shi (2011). International Journal of Data Warehousing and Mining (pp. 88-101).
www.igi-global.com/article/decision-rule-extraction-regularized-multiple/55080?camid=4v1a

Rainstorm Forecasting By Mining Heterogeneous Remote Sensed Datasets
www.igi-global.com/chapter/rainstorm-forecasting-mining-heterogeneous-remote/42370?camid=4v1a