Information Sciences 225 (2013) 1–17

Contents lists available at SciVerse ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Using sensitivity analysis and visualization techniques to open black box data mining models

Paulo Cortez [a,*], Mark J. Embrechts [b]

[a] Centro Algoritmi, Departamento de Sistemas de Informação, Universidade do Minho, Campus de Azurém, 4800-058 Guimarães, Portugal
[b] Department of Industrial and Systems Engineering Rensselaer Polytechnic Institute, CII 5129, Troy, NY 12180, USA

ARTICLE INFO

Article history:
Received 9 February 2012
Received in revised form 12 October 2012
Accepted 21 October 2012
Available online 17 November 2012

Keywords:
Sensitivity analysis
Visualization
Input importance
Supervised data mining
Regression
Classification

ABSTRACT

In this paper, we propose a new visualization approach based on a Sensitivity Analysis (SA) to extract human understandable knowledge from supervised learning black box data mining models, such as Neural Networks (NNs), Support Vector Machines (SVMs) and ensembles, including Random Forests (RFs). Five SA methods (three of which are purely new) and four measures of input importance (one novel) are presented. Also, the SA approach is adapted to handle discrete variables and to aggregate multiple sensitivity responses. Moreover, several visualizations for the SA results are introduced, such as input pair importance color matrix and variable effect characteristic surface. A wide range of experiments was performed in order to test the SA methods and measures by fitting four well-known models (NN, SVM, RF and decision trees) to synthetic datasets (five regression and five classification tasks). In addition, the visualization capabilities of the SA are demonstrated using four real-world datasets (e.g., bank direct marketing and white wine quality).

1. Introduction

Data Mining (DM) aims to extract useful knowledge from raw data. Interest in this field arose due to the advances of Information Technology and rapid growth of business and scientific databases [15]. These data hold valuable information such as trends and patterns, which can be used to improve decision making [30]. Two important DM tasks are classification and regression. Both tasks use a supervised learning paradigm, where the intention is to build a data-driven model that learns an unknown underlying function that maps several input variables to one output target.

Several learning models/algorithms are available for these tasks, each one with its own advantages. In a real-world setting, the value of a supervised DM model may depend on several factors, such as predictive capability, computational requirements and explanatory power. Often, it is important to have DM models with high predictive capabilities on unseen data. Computational effort and memory requirements are particularly relevant when dealing with vast datasets or real-time systems. This work focuses primarily on the explanatory power aspect, which relates to the possibility of extracting human understandable knowledge from the DM model. Such knowledge is important to determine if the obtained model makes sense to the domain experts and if it unveils potentially useful, interesting or novel information [15,4]. Increasing model interpretability allows for better understanding and trust of the DM results by the domain users [28] and this is particularly relevant in critical applications, such as control or medicine.

There is a wide range of "black box" supervised DM methods, which are capable of accurate predictions, but where obtained models are too complex to be easily understood by humans. This includes methods such as: Neural Networks (NNs)

* Corresponding author.
E-mail addresses: pcortez@dsi.uminho.pt (P. Cortez), embrem@rpi.edu (M.J. Embrechts).

0020-0255/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2012.10.039

(e.g., multilayer perceptrons and radial basis-functions) [18], Support Vector Machines (SVMs) and other kernel-based methods [10], and ensembles, including Random Forests (RFs) [2], where multiple models are combined to achieve a better predictive performance [11]. Recent examples of successful applications of these black box methods are: network intrusion detection using NN [16], wine quality prediction using SVM [7] and text sentiment classification (e.g., positive/negative movie-review identification) using ensembles of SVM and other DM methods [34].

To increase interpretability from black box DM models, there are two main strategies: extraction of rules and visualization techniques. The extraction of rules is the most popular solution [29,26,23]. However, such extraction is often based on a simplification of the model complexity, hence leading to rules that do not accurately represent the original model. For instance, a pedagogical technique was adopted in [27] within the intensive-care medicine domain to extract the relationships between the inputs and outputs of a NN classifier using a decision tree. While producing more understandable rules, decision trees discretize the classifier separating hyperplane, thus leading to information loss. Regarding the use of visualization techniques, the majority of these methods address aspects related to the multidimensionality of data and the use of visualization for black box DM models is more scarce [21]. Regarding the latter approach, some graphical methods were proposed, such as: Hinton and Bond diagrams for NN [9]; showing NN weights and classification uncertainty [31]; and improving the interpretability of kernel-based classification methods [5]. Yet, most of these graphical techniques are specific to a given learning method or DM task.

Our visualization approach to open DM models is based on a Sensitivity Analysis (SA), which is a simple method that performs a pure black box use of the fitted models by querying the fitted models with sensitivity samples and recording the obtained responses [25]. Thus, no information obtained during the fitting procedure is used, such as the gradient of the NN training or importance attributed to the splitting variable of a RF, allowing its universal application. In effect, while initially proposed for NN, SA can be used with virtually any supervised learning method, such as partial least squares [12] and SVM [7].

In [20], a computationally efficient one-dimensional SA (1D-SA) was proposed, where only one input is changed at the time, holding the remaining ones at their average values. Later, in [13] a two-dimensional SA (2D-SA) variant was presented. In both studies, only numerical inputs and regression tasks were modeled. Moreover, SA has been mostly used as a variable/feature selection method, where the method is used to select the least relevant feature that is deleted in each iteration of a backward selection [25,12,5,7].

The use of SA to open black box models was recognized in [20] but more explored in [13,21,8]. In [13], the proposed 2D-SA was used to show the effects of two input variables on the DM model, with the importance of these pair of inputs being measured by the simple output range measure. In [21], a genetic algorithm was used to search for interesting output responses related with one (2D plot) or two input (3D plot) variables. Yet, the study was focused on visualizing the individual predictions of an ensemble of models, where the intention was to check if the distinct individual predictions were similar, in conjunction with other criteria, such as the simpler output range measure. More recently, a Global SA (GSA) algorithm was presented in [8], capable of performing a F-dimensional SA for both regression and classification tasks, although with a high computational cost.

In this paper, we extend and improve our previous work [8], leading to a coherent SA framework capable of handing any black box supervised model, including ensembles, and applicable to both classification and regression tasks. The main contributions are:

(*i*) we present three novel and computationally efficient SA methods (DSA, MSA and CSA), comparing these with previous SA algorithms (1D-SA [20] and GSA [8]);
(*ii*) we propose a new SA measure of input importance (AAD), test it against three other measures, and present a more informative sensitivity measure pair for detecting 2D input relevance;
(*iii*) we adapt the SA methods and measures for handling discrete variables and classification tasks;
(*iv*) we propose novel functions for aggregating multiple sensitivity responses, including a 3-metric aggregation for 1D regression analysis and a fast aggregation strategy for input pair (2D) analysis;
(*v*) we present new synthetic datasets (four regression and five classification tasks) for evaluating input importance;
(*vi*) we present useful visualization plots for the SA results: input importance bars, color matrix, variable effect characteristic curve, surface and contour;
(*vii*) we explore three black box (NN, SVM and RF) and one white box (decision tree) models to test the SA capabilities and show examples of how SA can open the black box in four real-world tasks.

The paper is organized as follows. First, we present the SA approaches, visualization techniques, learning methods and datasets adopted in Section 2. Then, in Section 3 the proposed methods are tested in both synthetic and real-world datasets. Finally, conclusions are summarized in Section 4.

2. Materials and methods

2.1. Sensitivity methods

A supervised DM model is fit to a dataset, or training data, made up of N examples of M input variables and one output target (y). Let $\hat{y}$ denote the value predicted by the model for one example or data sample ($\mathbf{x}$) and let P be the function used to