Chapter 4

SYMBOLIC REGRESSION VIA GENETIC PROGRAMMING AS A DISCOVERY ENGINE: INSIGHTS ON OUTLIERS AND PROTOTYPES

Mark E. Kotanchek\textsuperscript{1}, Ekaterina Y. Vladislavleva\textsuperscript{2} and Guido F. Smits\textsuperscript{3}
\textsuperscript{1}Evolved Analytics L.L.C, Midland, MI, U.S.A.; \textsuperscript{2}University of Antwerpen, Antwerpen, Belgium; \textsuperscript{3}Dow Benelux B.V., Terneuzen, the Netherlands.

Abstract

In this chapter we illustrate a framework based on symbolic regression to generate and sharpen the questions about the nature of the underlying system and provide additional context and understanding based on multi-variate numeric data.

We emphasize the necessity to perform data modeling in a global approach, iteratively applying data analysis and adaptation, model building, and problem reduction procedures. We illustrate it for the problem of detecting outliers and extracting significant features from the CountryData\textsuperscript{1} – a data set of economic, political, social and geographic data collected. We present two complementary ways of extracting outliers from the data - the content-based and the model-based approach. The content-based approach studies the geometrical structure of the multi-variate data, and uses data-balancing algorithms to sort the data records in the order of decreasing typicalness, and identify the outliers as the least typical records before the modeling is applied to a data set. The model-based outlier detection approach uses symbolic regression via Pareto genetic programming (GP) to identify records which are systematically under- or over-predicted by diverse ensembles of (thousands of) global non-linear symbolic regression models.

Both approaches applied to the CountryData produce insights into outlier vs. prototypes division among world countries and about driving economic properties predicting gross domestic product (GDP) per capita.

Keywords: symbolic regression, data modeling, system identification, research assistant, discovery engine, outlier detection, outliers, prototypes, data balancing

\textsuperscript{1}http://reference.wolfram.com/mathematica/ref/CountryData.html
1. Introduction

The purpose of models is not to fit the data but to sharpen the questions.
–Samuel Karlin

Reality has a way of destroying beautiful theory. Thus, even though data modelers might construct beautiful algorithms, if the data does not agree with the implicit principles in that construct (e.g., system linearity, variable independence, variable significance, Gaussian additive noise) the house-of-cards comes tumbling down when it intersects with reality.

Pursuing data modeling as a main research direction, we have been building a framework based on symbolic regression to develop models which generate and sharpen the questions about what constitutes the underlying data-generating system. A useful framework helps us to understand what we know and do not know based on the data presented to us. We can begin to understand which data variables (or features, or attributes) are important and which are not, or whether we are missing some essential variables, because a reasonable prediction accuracy cannot be achieved. A good framework helps us to detect that some regions of the data space are either under- or over-represented.

Knowledge about these areas is essential for understanding the data. Data samples in over-represented areas can be flagged as prototypes, and possibly pruned for balancing the information content of samples over the data space. Samples in under-represented areas should be marked as outliers. They either represent measurement or computation errors, and should be removed from the modeling process, or on the contrary contain important nuggets of information about the system. In both cases the outliers are special, need to be treated with care during data interpretation and modeling, and always require human insight for the final verdict.

In this chapter we illustrate two sides of a holistic approach for understanding a multi-variate dataset from real-life - a collection of economic, political and geographic attributes gathered for 109 world countries. To understand and interpret the CountryData we present two approaches for outlier detection before and after the model development stage. The first is a content-based approach, which checks the spatial structure of the data. The second is a model-based approach. It uses symbolic regression to check the relationships among the attributes, and to extract the driving attributes for prediction of a characteristic economical feature of a country - the gross domestic product (GDP) per capita\(^2\). We apply two approaches to identify the special “outlier” countries:

- the countries, which are special, because they are spatially remote from the prototypic countries, and therefore are located in the under-

\(^2\)Gross domestic product per capita is the value of all final goods and services produced within a nation in a given year divided by the average population for the same year.