Chapter 1

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

Mark E. Kotanchek[1], Ekaterina Y. Vladislavleva[2] and Guido F. Smits[3]
[1]*Evolved Analytics L.L.C, Midland, MI, U.S.A.;* [2]*University of Antwerpen, Antwerpen, Belgium;* [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 the 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 [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

[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 data 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 charateristic 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-represented

[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

regions of the data space, and therefore require special treatment during modeling, analysis, measurement justification, etc. (content-based approach); and

- the countries, which are special because they possess an extraordinary GDPperCapita (extraordinary with respect to predictions of various ensembles of diverse symbolic regression models).

The first approach originates in our research on data balancing. It uses heuristic algorithms for weighting multi-variate input and input-output data, and for ordering the data in the order of decreasing importance - from outliers to prototypes (see (Vladislavleva, 2008)).

The second approach is model-based. It uses symbolic regression via genetic programming to generate ensembles of diverse regression models, which predict GDPperCapita attribute on the CountryData, and suggests outliers as points which consistently produce bad predictions on selected model ensembles.

Both approaches propose an interesting division of countries into "outliers" and "prototypes" without using any expert knowledge or interpretation of the CountryData. We believe that the hypothesis-generating aspect of symbolic regression enhanced with the insights from data balancing is essential for understanding multi-variate numeric data and data-generating system. It is also unique compared with other modeling methods, due to the transparency of explicit symbolic regression models.[3]

2. CountryData

The CountryData of Wolfram Research is a comprehensive collection of economic, geographic, social, and political data (224 attributes in total) over 237 world countries (taken from several credible sources like Encyclopaedia Britannica, United Nations Department of Economic and Social Affairs, United Nations Statistics Division, World Health Organization, and many others, see http://reference.wolfram.com/mathematica/note/CountryData- SourceInformation.html) We selected this data set because many attributes are highly correlated, so classic modeling methods alternative to symbolic regression would not be applicable; the dataset is of high dimensionality and heavily under-sampled (number of countries is approximately equal to the number of attributes; an average reader is aware of the economic positions of richest, poorest, and rapidly developing countries, which makes it easier to relate to the CountryData and interpret modeling results.

[3]The only other method with comparable power for discovery and insights is linear regression, but only in a situation where the underlying model structure is known, which is not the case in many real-life systems.

Our implementation of symbolic regression requires the data samples to be numeric, finite, and complete (no missing records), that's why we had to remove some countries and some attributes from the analysis, and got left with a list of 132 countries, and 128 attributes for them, including the GDPperCapita.

A challenge in our analysis is to reveal the relationship of the GDPperCapita of a country with other economic attributes, and to identify outlier countries with extraordinary GDPperCapita. To increase the chances of finding non-obvious relationships, we also excluded the attributes, which are explicitly related to GDP (we strive for insights, rather then for trivial relationships of the type GDPperCapita=GDP/TotalPopulation). All attributes other than GDPperCapita, containing the word GDP, or ValueAdded in their name were removed from the data set, e.g. AgriculturalValueAdded, ConstructionValueAdded, GDP, GDPAtParity, GDPPerCapita, GDPRealGrowth, IndustrialValueAdded, ManufacturingValueAdded, MiscellaneousValueAdded, NationalIncome, TradeValueAdded, TransportationValueAdded, ValueAdded.

The remaining attributes for the analysis are:
CountryIndex, AdultPopulation, Airports, AMRadioStations, AnnualBirths, AnnualDeaths, ArableLandArea, ArableLandFraction, Area, BirthRateFraction, BoundaryLength, CallingCode, CellularPhones, ChildPopulation, CoastlineLength, CropsLandArea, CropsLandFraction, DeathRateFraction, EconomicAid, ElderlyPopulation, ElectricityConsumption, ElectricityExports, ElectricityImports, ElectricityProduction, ExchangeRate, ExportValue, ExternalDebt, FemaleAdultPopulation, FemaleChildPopulation, FemaleElderlyPopulation, FemaleInfantMortalityFraction, FemaleLifeExpectancy, FemaleLiteracyFraction, FemaleMedianAge, FemalePopulation, FixedInvestment, FMRadioStations, GovernmentConsumption, GovernmentExpenditures, GovernmentReceipts, GovernmentSurplus, GrossInvestment, HighestElevation, HouseholdConsumption, ImportValue, InfantMortalityFraction, InflationRate, InternetHosts, InternetUsers, InventoryChange, IrrigatedLandArea, IrrigatedLandFraction, LaborForce, LandArea, LifeExpectancy, LiteracyFraction, LowestElevation, MaleAdultPopulation, MaleChildPopulation, MaleElderlyPopulation, MaleInfantMortalityFraction, MaleLifeExpectancy, MaleLiteracyFraction, MaleMedianAge, MalePopulation, MedianAge, MigrationRateFraction, MilitaryAgeMales, MilitaryExpenditureFraction, MilitaryFitMales, NaturalGasConsumption, NaturalGasExports, NaturalGasImports, NaturalGasProduction, NaturalGasReserves, OilConsumption, OilExports, OilImports, OilProduction, PavedAirports, PavedRoadLength, PhoneLines, Population, PopulationGrowth, PriceIndex, RadioStations, RoadLength, ShortWaveRadioStations, TelevisionStations, TotalConsumption, TotalFertilityRate, UnemploymentFraction, UNNumber, WaterArea, ExpenditureFractions• { ExportValue, FixedInvestment, GovernmentConsumption, GrossInvestment, HouseholdConsumption, ImportValue, InventoryChange }, TotalConsumption, PavedAirportLengths • { 3000To5000Feet, 5000To8000Feet, 8000To10000Feet, Over10000Feet, Total, Under3000Feet }.

This chapter focuses on outlier detection, so our goal in the analysis of the CountryData is to extract the countries out of 132 available, which are special, i.e. they deviate from the prototypic countries with 'normal' economic indicators. We are striving to develop a research assisting framework for data analysis, and thus our 'outlier' detecting techniques should suggest 'outlier' candidates to the domain expert, but should not use any expert knowledge during the identification process. The expert is the one to decide what to do with suggested outliers, and he or she is the one to gain additional insights from these. The data analysis system is just an enabling technology that triggers

the expert to ask a new question, and learn something new about the data-generating system.

3. Data balancing as an insightful pre-processing step and content-based outlier detection

Data weighting for detecting under-represented regions of the data space

Pre-processing and scrutinizing data is a crucial first step of the learning process. Constructing bivariate plots of all variable pairs and computing a correlation matrix of data can sometimes reveal strong linear dependencies among variables of interest. This can allow breaking the data down into sets of smaller dimensionality, which are easier to explore visually, and to reveal outliers. However, when the data is of high dimensionality and very sparse (we have 132 records and 128 variables in the CountryData), visual exploration of bivariate plots of data for potential outliers is, first, time-consuming, and, second, is risky in terms of being deceptive.

In this section we describe a more structured and automated approach of exploring the geometric structure of data. It does not make any assumptions about the underlying relationships among data variables, and identifies the records, which are spatially remote from other records in the data space. We refer to it as to Data Balancing, since the approach belongs to a suite of techniques for analysis, adaptation and modeling of imbalanced data, see (Vladislavleva, 2008).

In (Vladislavleva, 2008) several algorithms for weighting and balancing multi-variate input- and input-output data are presented. Data weighting assigns weights to data records, and the weight is interpreted as a measure of relative importance (information content) of that data record. Information content is connected to the sparsity of the neighborhood of a data sample. It can reflect the proximity of a sample to its k nearest or nearest-in-the-input space neighbors, the surrounding of a sample by k nearest or nearest-in-the-input-space neighbors, or the local deviation from a hyper-plane approximating k nearest-in-the input space neighbors. The first two weights are introduced in (Harmeling et al., 2006) for unlabeled data, and are further extended to include input-response data and use a particular fractional distance metric.

Due to space limitations of this chapter we will give definition for one weighting functional only - the surrounding weight.

By input-output data we mean a set $\mathcal{M} = \{M_1, \ldots, M_N\}$ of N points in a $(d+1)$-dimensional space $\mathbb{R}^{d+1}$. Point M_i has coordinates $(x_1^i, x_2^i, \ldots, x_d^i, y^i) \in \mathbb{R}^{d+1}$, $i = \overline{1, N}$, with y^i corresponding to the response value at the input point $P_i = (x_1^i, x_2^i, \ldots, x_d^i) \in X \subset \mathbb{R}^d$. We say that the input-output point M_i represents the input point P_i, since the projection of M_i on the input space

$X \subset \mathbb{R}^d$ is exactly P_i. The set of all input points is denoted as $\mathcal{P}$, the vector of outputs as $Y = (y^1, \ldots, y^N)^T$.

By $\{n_1(P_i, \mathcal{P}), n_2(P_i, \mathcal{P}), \ldots, n_k(P_i, \mathcal{P})\} \in \mathcal{P}$ we denote k nearest neighbors of the point $P_i \in \mathcal{P}$ in metric L_2 or $L_{1/d}$.

For unlabeled data the surrounding weight is defined as the length of the sum of vectors connecting a sample with its k nearest neighbors (averaged over k):

$$\sigma(i, \mathcal{P}, k) = \left|\left| \frac{1}{k} \sum_{j=1}^{k} (P_i - n_j(P_i, \mathcal{P})) \right|\right|, \tag{1.1}$$

where the $n_j(P_i, \mathcal{P})$ is the j-th nearest neighbor of the point P_i from the set $\mathcal{P}$ in the norm $\| \cdot \|_{1/d}$ or $\| \cdot \|_2$.

For labeled data, the surrounding weight is defined as the length of the sum of vectors connecting a sample with its k nearest-in-the-input space neighbors:

$$\sigma(i, \mathcal{M}, \mathcal{P}, k) = \left|\left| \frac{1}{k} \sum_{j=1}^{k} (M_i - \bar{n}_j(M_i, \mathcal{M}, \mathcal{P}) \right|\right|, \tag{1.2}$$

where $\bar{n}_j(M_i, \mathcal{M}, \mathcal{P})$ is the j-th nearest-in-the-input-space neighbor of point M_i in the set $\mathcal{M}$. This means that the projection of $\bar{n}_j(M_i, \mathcal{M}, \mathcal{P})$ onto the input space X is $n_j(P_i, \mathcal{P})$.

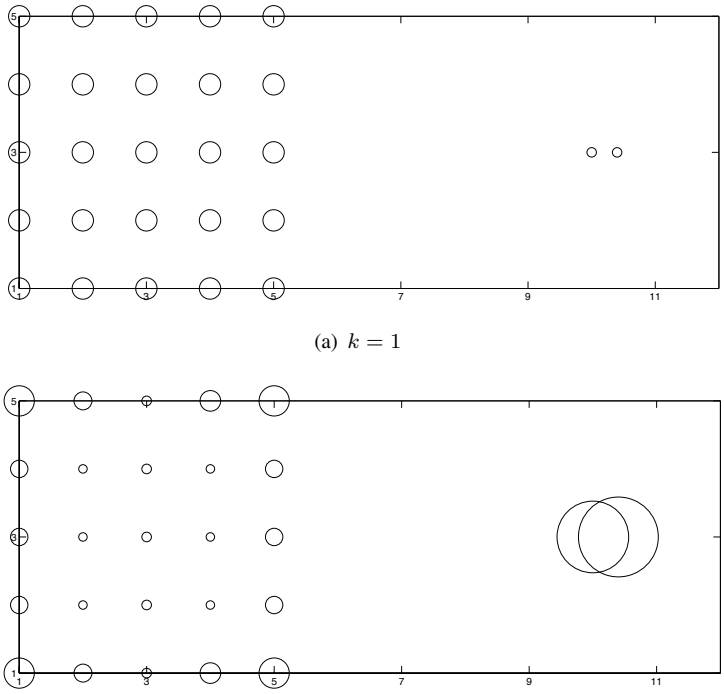
The neighborhood definition changes to reflect the fact that labeled data is assumed to belong to the response-surface, and in the input-output data space the notion of closeness to the closest neighbors can be deceiving (points, which are the closest in the input-response space may not be the closest on the response surface).

Data balancing for constructing of smaller subsets with similar information content

The neighborhood size k dictates the scale in the perception of the data - small neighborhood size suggests local analysis of data, while big neighborhood size implies a global view on data. If the weights are computed for $k = 1$ - they reflect the local importance of a point - only relative to the nearest neighbor. Thus, points with high weights are the ones, which are remote even from their nearest and nearest-in-the-input space neighbors. Therefore, these points are located in remote clusters of size one, and hence can be interpreted as 'outliers'.

This approach of identifying remote clusters of points as outliers stops to work if the size of the clusters is bigger that the neighborhood size k used in weight computation.

For example, the points which are located in remote clusters of size $k + 1$, will all have small weights computed for the neighborhood k, because they are



(a) $k = 1$

(b) $k = 5$

Figure 1-1. Straightforward weight computation in one pass cannot discover remote clusters of data if the neighborhood size is smaller than the size of the cluster.

close to it's k neighbors in the cluster. Small weights will not alarm about the remoteness of such a cluster relative to the entire data set. To discover, that the cluster is remote, we'd need to compute weights of points with neighborhoods larger than the cluster size.

In Figure 1-1 we illustrate the problem of detecting remote data clusters with one-pass weighting. Figure 1-1 shows the unlabeled data set of 27 points, two of which are located in a sparse region of data space. In plot 3.0 we indicate the surrounding weights of 27 points computed with 1 nearest neighbor (weights correspond to the radii of the circles with centers in data points). The weights of the remote points are very small in this case relative to the weights of other points in the data set. A simple observation of the sorted weights profile does not provide insights into the fact that the two points are outliers.

In plot 3.0 we show the surrounding weights computed with the neighborhood size of $k = 5$. Since the k is greater than the size of the remote cluster - the large weights of the points in the cluster do reflect the discrepancy in the spatial location of cluster points relative to other points in the data set.

To still be able to identify the records belonging to regions which are globally remote from prototypic samples, but locally densely populated, we use a heuristic algorithm which iteratively eliminates records with the smallest weight from the data set, and recomputes the weights for points which had the eliminated record among the k nearest, or nearest-in-the-input-space neighbors. This procedure, called a Simple Multi-dimensional Iterative Technique for Sub-sampling (SMITS), gradually prunes the data set starting from the most densely populated regions, removes records with the smallest obtained weight from the data set, and ranking these points by the order, in which they are eliminated. After the record is eliminated from the data set, the weights of the points that had the eliminated point among k nearest or nearest-in-the-input-space neighbors must be re-evaluated at each iteration step. At the end of the elimination procedure only k points remain in the data set, and those points are randomly ranked by indexes $N - k + 1, \ldots, N$.

We interpret the elimination rank as a measure of the global relative importance of a data point. Points, which are representatives of the dense clusters will be eliminated the last. The weights of these points gradually increase as their neighbors get eliminated. At the moment when a point gets eliminated (which happens if the current weight of this point is the smallest among all points left) - the weight of the point represents the cumulative weight of the cluster, in which the point was originally located. For this reason archiving the weights of eliminated points at each elimination step - will get us the ordering of data records from prototypes (located in well- or over-represented regions) to the "back-bone" points (forming a space-filling support structure of the data set).

The geometrical outliers, or representatives of clusters of outliers (as in the example with three points) will never be eliminated before the prototypic points which have a smaller weight. Therefore the outliers, or representatives of the clusters of outliers will stay in the data set till the last stage of the elimination process[4].

The SMITS procedure defines an order of the data records, which can be used to partition the data set into nested subsets of increasing size (when the eliminated records are added one by one to a subset of k records). If during the elimination process we archive a weight of each eliminated point, we can compute the cumulative sum of these weights for each elimination step in the order opposite to the order of elimination. If the weights are normalized by the total sum of the weights of the data set they will sum up to the number of records N. After normalization the resulting cumulative sum of eliminated

[4]the elimination stops when there are k records left n the data set, since the surrounding weight weight relative to k nearest or nearest-in-the-input space neighbors will not be defined for less than k records.

weights can be interpreted as a cumulative information content of the data set ranked with the SMITS procedure.

We illustrate the procedures of data weighting and cumulative information content calculation on the CountryData.

Scaling the CountryData

If all attributes are equally important for the modeler, we suggest scaling the data to a standard range before weighting and balancing it. But rather than scaling the ranges of all attributes to a particular range, e.g. $[0, 1]$, we advise mapping the 10th and 90th percentiles of the attribute ranges to the ends of selected interval. This decreases the sensitivity of the scaled results to outliers in the records.

Insights for CountryData

Weighting the data: Since we are interested in the outlier-countries with atypical GDPperCapita we turn our data into input-response data, with 108 input attributes, and one response attribute - GDPperCapita. We weight the data with the surrounding weight and one nearest-in-the-input-space neighbor. By definition of the surrounding weight, the countries located in the very sparse regions of the data space will get the highest surrounding weights. In Figure 1-2 we plot the sorted surrounding weights computed for one neighbor (the nearest neighbors are determined in the input space, and distances are computed in the 109 dimensional input-output space). The weights are normalized, so they sum up to the number of countries. From the plot we can infer that the weights of five countries are radically different from the rest. These countries with corresponding single-pass weights are UnitedStates (weight 18.1), Russia (weight 7.7), Canada (weight 6.2), India (weight 5.9), and China (weight 5.9). Other countries all have weights smaller than 3.6 (the weight of Japan), and can be considered as prototypes, since they are located in the better proximity to their nearest-in-the-input space neighbors[5].

Since the Euclidean distance $L_2(p,q) = \left(\sum_{j=1}^{d}(p_i - q_i)^2\right)^{1/2}$, $p, q \in \mathbf{R}^d$ was shown to fail in giving a meaningful notion of proximity in a high-dimensional space, we suggest using a fractional distance metric $L_{1/d}$ in a d-dimensional space, when d is large:

$$dist_{1/d}(\mathbf{u}, \mathbf{v}) = \left(\sum_{i=1}^{d} |u^i - v^i|^{1/d}\right)^d ,$$

[5]The results of the weighting almost do not change in this example compared with weighting of ContryData without specifying GDPperCapita as a response variable. Only the rankings of 7 countries in the mid-weights change slightly for a different definition of nearest neighbors.

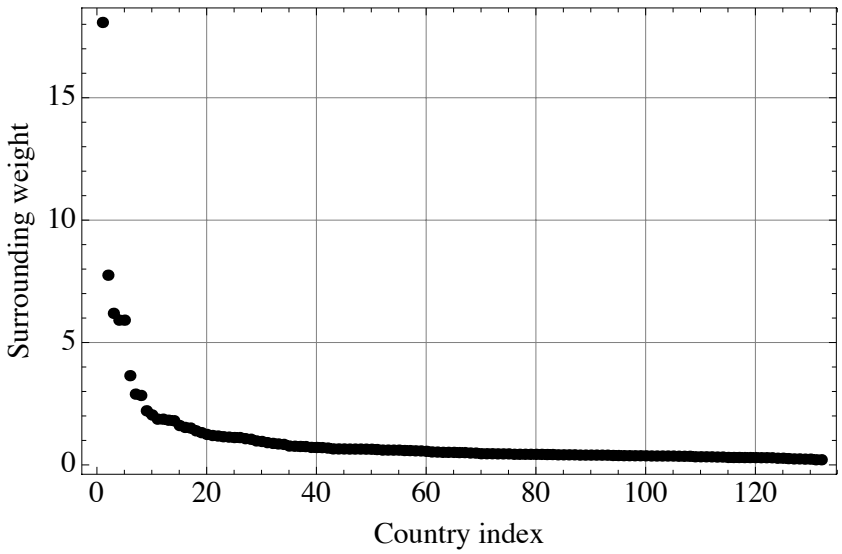


Figure 1-2. Exploration of the weights of data records gives insights into the remote outlier records located in sparse under-represented regions of the data space.

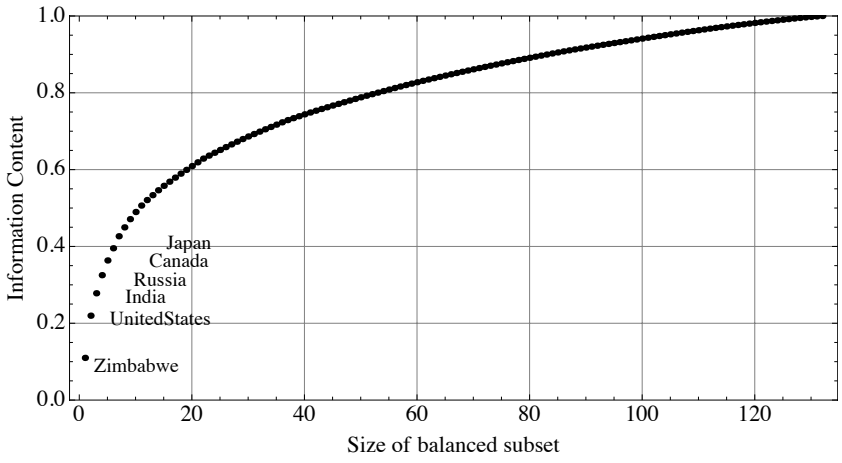
$$L_{1/d} : \|\mathbf{u}\|_{1/d} = \left(\sum_{i=1}^{d} |u^i|^{1/d} \right)^d .$$

A fractional distance metric in a space of high dimensionality can scale better, see (Aggarwal et al., 2001). See (Francois et al., 2007) for the detailed discussion on the relevance of using fractional distances with respect to the distance concentration phenomenon.

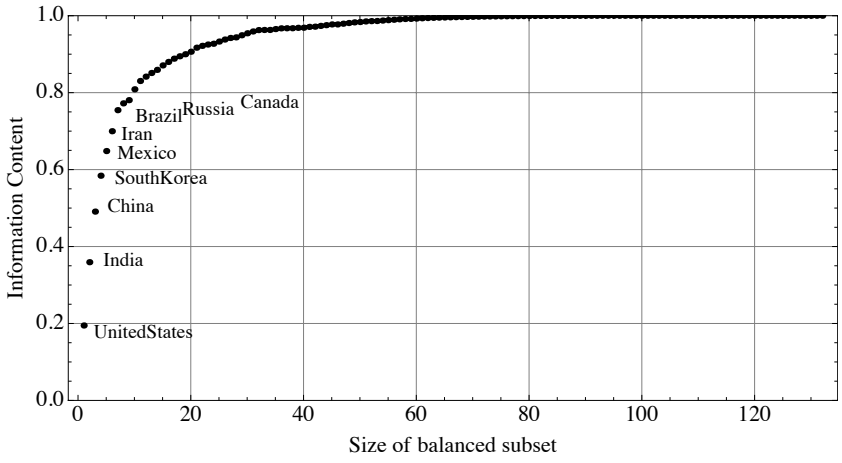
Balancing the data: Now we apply a balancing heuristic (the SMITS algorithm) to order the CountryData with GDPperCapita as a response variable in the order of decreasing importance. We again use one nearest-in-the input space neighbor, and two distance measures.

In Figure 1-3 we plot the cumulative information contents (CIC) of the CountryData ranked with surrounding weight via the SMITS procedure using two distance metrics. A value on a curve at point m is by definition a cumulative elimination weight of the first m samples in the balanced subset. It can be interpreted as a fraction of the information about the data contained in the first m samples of the ranked data set, $m = 1 : N$.

We can observe that the shape of the cumulative information content and also the SMITS-based ranking of countries changes, if the distance metric changes. From the plot 1-3(b) we can observe that the first nine to ten countries in the CountryData subset, balanced with the factional distance metric, are representing 80% of the information content (i.e. have the cumulative elimination



(a) Euclidean distance metric. $k = 1$



(b) Fractional (1/108)-distance metric. $k = 1$

Figure 1-3. Cumulative information content of the balanced data set can indicate compressibility of the set, and the fractions of outliers and prototypes.

weight of 0.8). The back-bone countries, i.e. outliers and the most representative prototypes are contained in this subset of countries. Since the own elimination weights of the countries also indicate the 'cumulative' weigh of the dense data clusters, that these 'back-bone' points are representing - the relative difference of these weights (i.e. implicitly the shape of the cumulative information content curve) provides insights into the likelihood of the mentioned clusters to be outliers.

So, the data balancing with euclidean distance and neighborhood of $k = 1$ produces the ranking of countries with the following elimination weights[6]: Zimbabwe (14.1), UnitedStates (14.1), India (7.7), Russia (6.2), Canada (5.6), Japan (4.2), China (4.1), UnitedKingdom (3.1), Brazil (2.8), Italy (2.4), Indonesia (2.2), and other countries, whose elimination weights quickly decrease below 2. Observing the differences we can hypothesize, that the seven to ten countries represent data clusters, which are likely to be outliers.

The data balancing with a fractional $L_1/109$ distance and $k = 1$ produces the following elimination weights: UnitedStates (38.3), India (38.3), China (12.9), SouthKorea (3.9), Mexico (6.1), Iran (5.4), Brazil (5.3), Russia (4.4), Canada (2.9), and other countries, whose weights quickly decrease in values below 2.

We provide the interested reader with the rankings of the subset of 30 countries obtained with weighting and balancing with different distance metrics in Table 1-1.

Identifying outliers among attributes by balancing the transposed data matrix: We note that the data balancing algorithm can also be applied to identify the 'back-bone' attributes of the data set. By applying the SMITS algorithm to the transposed data matrix of the CountryData, and treating attributes as records - we can get interesting insights into attributes, which are representative of the entire set of 109 attributes. By using a fractional distance metric $L_{1/132}$ and balancing the transposed data matrix (of the size 109×132) with SMITS for the neighborhood size of $k = 1$ we obtain a content-based ranking of attributes and their elimination weights. Because of space limitations we cannot give a full ranking of attributes, but would like to share the top ones with the reader. The *blind* data balancing algorithm applied to the scaled (transposed) data matrix (without any *a priori* information about the importance, or preferences for any attributes) discovers the following top four attributes for the CountryData: GDPPerCapita, ExpenditureFractions•TotalConsumption, CallingCode, and MaleLiteracyFraction!

In the next section we use symbolic regression to discover relationships of the GDPperCapita with other attributes of the CountryData. After creating

[6]For comparison purposes all elimination weights are normalized to sum up to the total number of records (132 in this case).

Country	**Weight rank**	SMITS **Euclidean**	SMITS **Fractional**
UnitedStates	1	2	1
Russia	2	4	8
Canada	3	5	9
India	4	3	7
China	5	7	3
Japan	6	6	29
Brazil	7	9	7
UnitedKingdom	8	8	40
Indonesia	9	11	57
Germany	10	12	65
Italy	11	10	12
France	12	29	14
Iran	13	22	6
Qatar	14	14	67
SaudiArabia	15	15	79
Australia	16	20	64
Lesotho	17	16	110
Mexico	18	25	5
Turkmenistan	19	27	34
Vietnam	20	13	33
Norway	21	18	24
Nigeria	22	20	72
Bangladesh	23	32	10
Spain	24	24	28
Paraguay	25	26	19
Peru	26	21	31
Algeria	27	31	25
Philippines	28	28	37
Turkey	29	23	53
SouthKorea	30	36	4

Table 1-1. Content-based rankings of 30 countries of the CountryData for $k = 1$. The first column represents the ranking obtained by sorting the one-pass surrounding weights computed with Euclidean distance. The second and the third columns represent ranking obtained with data balancing via SMITS algorithm with Euclidean distance and fractional $1/109$-distance respectively. Observe that changing the distance generates different 'back-bone' countries as representatives of the dense clusters of countries.

hundreds of explicit regression models we define outliers as countries, whose GDPperCapita is consistently under or over-predicted by the constructed models.

4. Symbolic regression and model-based outlier detection

Symbolic regression as a modeling engine

Symbolic regression via genetic programming (GP) is a non-parametric non-linear regression technique that looks for an appropriate model structure and model parameters (as opposed to classic regression that assumes a certain model structure and estimates the optimal parameters). Symbolic regression is an attractive modeling engine because it mitigates the need to make a cascade of simplifying assumptions about the system and, instead, allows the data to define the appropriate model forms (provided it is used with intelligent and rigorous complexity control!). We use a particular flavour of genetic programming, with a multi-objective selection operator, which favors high prediction accuracy and low model complexity in models during selection process. Propagation rights are distributed among individuals satisfying some conditions on Pareto-optimality in the objective space of accuracy vs. complexity. We refer to this methodology as to symbolic regression via Pareto genetic programming (Pareto GP), see (Smits and Kotanchek, 2004),(Kotanchek et al., 2006).

The strongest capabilities of modeling with symbolic regression via Pareto GP (and with other GP flavors with incorporated complexity control, fitness inheritance for successful variables, ensemble-based predictions) are:

- automatic (and robust) selection of significant variables related to a response variable (see e.g. (Smits et al., 2005));
- automatic generation of diverse model structures describing the relationship between the response and significant input variables;
- automatic generation of ensembles of diverse prediction models (all of which are global learners, but are constrained to be diverse with respect to complexity and uncorrelated w.r.t. prediction error), see (Kotanchek et al., 2007);
- automatic "outlier" identification (where outliers are defined as samples, which persistently imply worse prediction errors compared with the average (or prototypic) data samples on selected ensembles of diverse regression models).

With symbolic regression we can discover simplifying variable transformations, that make subsequent modeling cycles more efficient, accurate, and easier to interpret and might potentially reveal something about the underlying

physical system. With Symbolic regression via Pareto GP we can identify optimal trade-offs between competing modeling structures. We can begin to understand what is the dimensionality of the space that is sufficient to describe the system. We can exploit the multitude of solutions of competing accuracy and complexities generated by symbolic regression to our advantage: ensembles of diverse models can always be used to generate an estimate of prediction trustworthiness that guides the user in his exploration of the new areas of the data space.

The hypothesis generating aspects of symbolic regression embedded into a global data modeling framework turn it into a valuable research assistant. Rather than eliminating the modeler from the modeling cycle, it frees more thinking time for a modeler, triggers new ideas and reveals the flaws in existing levels of understanding.

Model ensembles and suggested "outliers" on CountryData

We executed 150 independent runs of 10 minutes each for modeling GDPperCapita via other 108 attributes of the CountryData. All runs used the default settings Pareto GP evolutionary strategy - two-objective selection with archiving, 300 population individuals, 100 archive members, 95%crossover, 5% mutation rates, and standard basis functions - multiplication, subtraction, product, division, inverse, negation, square root, and square (maximal arity of non-unary operators limited to 4). Objective functions for pareto-based model selection were normalized sum of squared errors (scaled to the interval $[0, 1]$, with zero corresponding to the perfect error) and expressional complexity, computed as the total sum of nodes in all subtrees of a model tree. In both objectives smaller values are preferred.

Variable Selection

The GP runs generated more than 18000 regression models. We selected 3914 'interesting' models with expressional complexity of at most 150, and normalized prediction error of at most 0.3. These models were inspected for variable presence to identify the driving attributes related to the GDPperCapita. We plot the variable presence map in Figure 1-4. The following (correlated) attributes were present in the 3914 interesting models: TotalConsumption (in 86% models), FixedInvestment (68%), GovernmentReceipts (59%), Population (45%), GrossInvestment (42%), AdultPopulation (28%), FemalePopulation (21%), HouseholdConsumption (21%), FemaleAdultPopulation (20%), GovernmentSurplus(16%), GovernmentConsumption (12%), ExportValue (12%), ImportValue(8%), GovernmentExpenditures (6%). It is interesting to observe that the Population-related variable is clearly a driving attribute for predicting the GDPperCapita, and attributes of consumption, in-

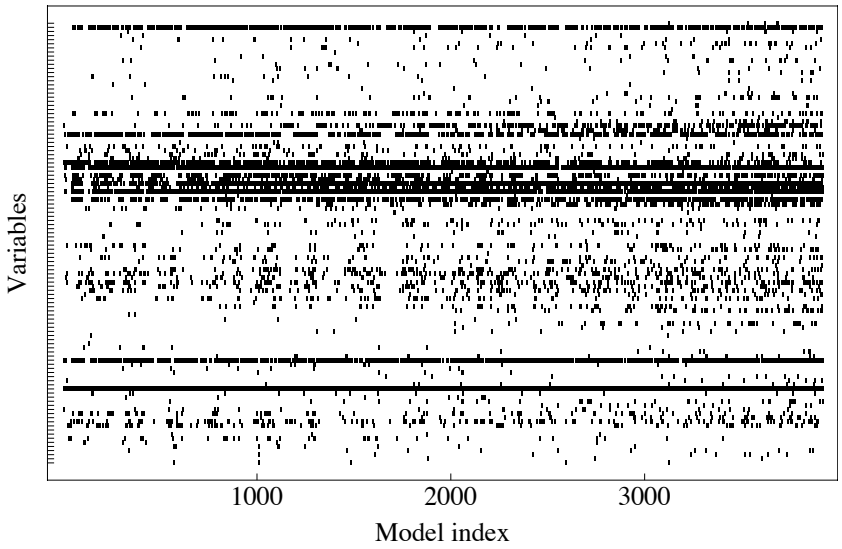
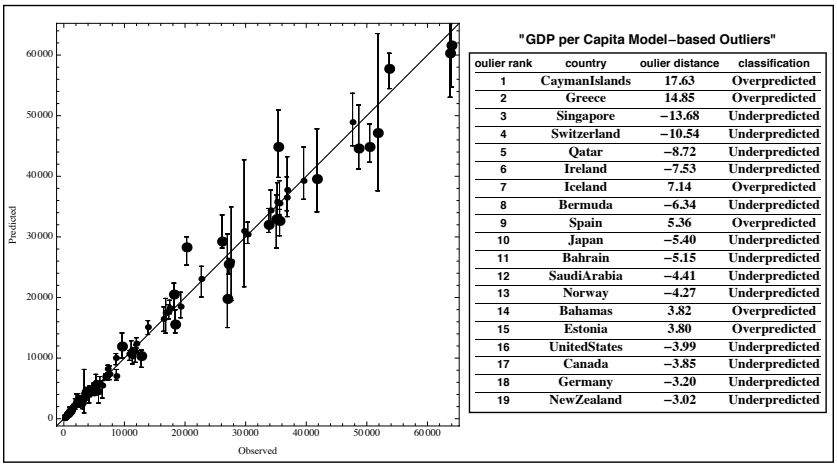


Figure 1-4. Symbolic regression can be used to discover significant attributes related to GDPperCapita in robust and reproducible fashion.

vestment, import, and export have a stronger relation to GDPperCapita (according to symbolic regression results) than NaturalGasProduction, or RoadLength, etc. The attributes, which are not listed above appeared in less than 5% of interesting models.

Ensemble construction and outlier identification

We pruned the set of interesting models further to automatically select 927 candidate models, from which an ensemble of 17 diverse accurate and parsimonious models was created. In the left plot of Figure 1-5 we plot the prediction of this ensemble. Based on deviation of predictions of ensemble models, we define a measure of ensemble disagreement for each point of the data space, where the prediction is computed. In this way solutions of symbolic regression become trustworthy - with each prediction a confidence interval is supplied, see (Kotanchek et al., 2007) for more details. As soon as an ensemble is created, we can analyze data records with respect to systematic errors in predictions of ensemble models. Records, which are consistently under- or over-predicted by an ensemble are candidates for outliers. The top 19 countries deviating from ensemble predictions are shown in the right of Figure 1-5. Relative to the constructed models such countries as CaymanIslands, Greece, Iceland, Spain, Bahamas, and Estonia are underperforming, and should have shown a higher GDPperCapita than actual. The countries like Singapore, Switzerland, Qatar, Ireland, Bermuda, Japan, Bahrain, and others are on the contrary over-



"GDP per Capita Model−based Outliers"

oulier rank	country	oulier distance	classification
1	CaymanIslands	17.63	Overpredicted
2	Greece	14.85	Overpredicted
3	Singapore	−13.68	Underpredicted
4	Switzerland	−10.54	Underpredicted
5	Qatar	−8.72	Underpredicted
6	Ireland	−7.53	Underpredicted
7	Iceland	7.14	Overpredicted
8	Bermuda	−6.34	Underpredicted
9	Spain	5.36	Overpredicted
10	Japan	−5.40	Underpredicted
11	Bahrain	−5.15	Underpredicted
12	SaudiArabia	−4.41	Underpredicted
13	Norway	−4.27	Underpredicted
14	Bahamas	3.82	Overpredicted
15	Estonia	3.80	Overpredicted
16	UnitedStates	−3.99	Underpredicted
17	Canada	−3.85	Underpredicted
18	Germany	−3.20	Underpredicted
19	NewZealand	−3.02	Underpredicted

Figure 1-5. Ensembles of symbolic regression models can be used to discover records, which systematically disagree with predictions of individual ensemble models.

performing, which might imply that some other hidden attributes not included in our current list are contributing to an increased GDPperCapita. This in itself is an interesting and very relevant result of this type of data analysis using symbolic regression.

In such an interpretation the outliers are defined as records, which disagree with produced models. Such an approach is viable only if produced models are reliable, and provide plausible descriptions of the system or process This requires a domain expert to take decisions about reliability of produced models. There are, however, two factors, which contribute to justification of the approach and to convincing the expert to exploit symbolic regression as a research assistant:

- The fact that thousands of diverse regression models are created during the modeling process without external assumptions about the model structure implies, the models in the ensemble have maximal predictive power and the highest reliability warranted by the data.

- The fact that the ensemble of multiple symbolic regression models is used to decide whether a record is an outlier or not makes this approach robust (the more models from a reliable and diverse collection identify a record as an outlier, the more likely this is actually true).

5. Conclusions and Recommendations

Outlier detection for nonlinear systems with lots of input variables is hard to achieve using conventional methods, and is dangerous to perform if samples

are defined as outliers w.r.t. a model, and the model is " guessed" incorrectly. An outlier is either the most important nugget in the data set or something which should be removed from the modeling process to avoid distorting the results. Interpretation of an outlier always requires human insight of a domain expert.

We have shown two approaches for (automatic) outlier detection both before and after the model development process. It is automatic in a sense that no assumptions or guesses should be made about the model or about the trustworthiness of the data records up until the final stage, where the outliers are identified. At that point the user or the modeler need to make a decision about an outlier's destiny - it is a "jewel or a junk". We believe that both approaches are viable and need to be used in combination to fully exploit the power of symbolic regression as a discovery engine.

References

Aggarwal, Charu C., Hinneburg, Alexander, and Keim, Daniel A. (2001). On the surprising behavior of distance metrics in high dimensional space. *Lecture Notes in Computer Science*, 1973:420–434.

Francois, Damien, Wertz, Vincent, and Verleysen, Michel (2007). The concentration of fractional distances. *IEEE Trans. on Knowledge and Data Engineering*, 19(7):873–886.

Harmeling, Stefan, Dornhege, Guido, Tax, David, Meinecke, Frank, and Muller, Klaus-Robert (2006). From outliers to prototypes: Ordering data. *Neurocomputing*, 69(13-15):1608–1618.

Kotanchek, Mark, Smits, Guido, and Vladislavleva, Ekaterina (2006). Pursuing the pareto paradigm tournaments, algorithm variations & ordinal optimization. In Riolo, Rick L., Soule, Terence, and Worzel, Bill, editors, *Genetic Programming Theory and Practice IV*, volume 5 of *Genetic and Evolutionary Computation*, chapter 12, pages 167–186. Springer, Ann Arbor.

Kotanchek, Mark, Smits, Guido, and Vladislavleva, Ekaterina (2007). Trustable symoblic regression models. In Riolo, Rick L., Soule, Terence, and Worzel, Bill, editors, *Genetic Programming Theory and Practice V*, Genetic and Evolutionary Computation, chapter 12, pages 203–222. Springer, Ann Arbor.

Smits, Guido, Kordon, Arthur, Vladislavleva, Katherine, Jordaan, Elsa, and Kotanchek, Mark (2005). Variable selection in industrial datasets using pareto genetic programming. In Yu, Tina, Riolo, Rick L., and Worzel, Bill, editors, *Genetic Programming Theory and Practice III*, volume 9 of *Genetic Programming*, chapter 6, pages 79–92. Springer, Ann Arbor.

Smits, Guido and Kotanchek, Mark (2004). Pareto-front exploitation in symbolic regression. In O'Reilly, Una-May, Yu, Tina, Riolo, Rick L., and Worzel,

Bill, editors, *Genetic Programming Theory and Practice II*, chapter 17, pages 283–299. Springer, Ann Arbor.

Vladislavleva, Ekaterina (2008). *Model-based Problem Solving through Symbolic Regression via Pareto Genetic Programming*. PhD thesis, Tilburg University, Tilburg, the Netherlands.