Liu et al. BMC Systems Biology 2014, 8(Suppl 3):S5
http://www.biomedcentral.com/1752-0509/8/S3/S5

BMC Systems Biology

RESEARCH Open Access

Learning accurate and interpretable models based on regularized random forests regression

Sheng Liu[1], Shamitha Dissanayake[2], Sanjay Patel[3], Xin Dang[4], Todd Mlsna[2], Yixin Chen[1*], Dawn Wilkins[1]

From IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2013)
Shanghai, China. 18-21 December 2013

Abstract

Background: Many biology related research works combine data from multiple sources in an effort to understand the underlying problems. It is important to find and interpret the most important information from these sources. Thus it will be beneficial to have an effective algorithm that can simultaneously extract decision rules and select critical features for good interpretation while preserving the prediction performance.

Methods: In this study, we focus on regression problems for biological data where target outcomes are continuous. In general, models constructed from linear regression approaches are relatively easy to interpret. However, many practical biological applications are nonlinear in essence where we can hardly find a direct linear relationship between input and output. Nonlinear regression techniques can reveal nonlinear relationship of data, but are generally hard for human to interpret. We propose a rule based regression algorithm that uses 1-norm regularized random forests. The proposed approach simultaneously extracts a small number of rules from generated random forests and eliminates unimportant features.

Results: We tested the approach on some biological data sets. The proposed approach is able to construct a significantly smaller set of regression rules using a subset of attributes while achieving prediction performance comparable to that of random forests regression.

Conclusion: It demonstrates high potential in aiding prediction and interpretation of nonlinear relationships of the subject being studied.

Background

In many real applications, it is vital to have an interpretable model (e.g., relevant features and predictive rules) and high performance prediction at the same time to understand the underlying problem well. Some of the state-of-the-art algorithms like Support Vector Machines (SVM) [1], Artificial Neural Network (ANN) [2], and Random Forests (RF) [3], generally predict the outcome with high accuracy. But other than accuracy, it is hard to interpret the models built since they either are "black box" model, or include so many decision rules that human cannot explain them clearly. On the other hand, some algorithms, especially those based on decision trees, are easy to interpret. However, the prediction performance is usually low compared to SVM, ANN, or RF. See Figure 1 for an illustration regarding the interpretability-prediction performance space. Basically, to help explain the generated model, it is desirable to have an algorithm that falls on the region C. Finding a right tradeoff between prediction performance and model interpretability is thus important.

Decision trees use a tree structure to represent a partition of the space. From the root node to each leaf node of a decision tree, we can consider it as a decision rule. Decision rule based algorithms are well known for their capability of shedding light on the decision process in addition to making a prediction. Another factor affecting the interpretation of model generated from data is feature selection. In general, fewer features involved in the model will make it less complex and

* Correspondence: ychen@cs.olemiss.edu
[1]Department of Computer and Information Science, University of Mississippi, Weir Hall 201, 38677, University, MS, U.S.A
Full list of author information is available at the end of the article

BioMed Central

© 2014 Liu et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

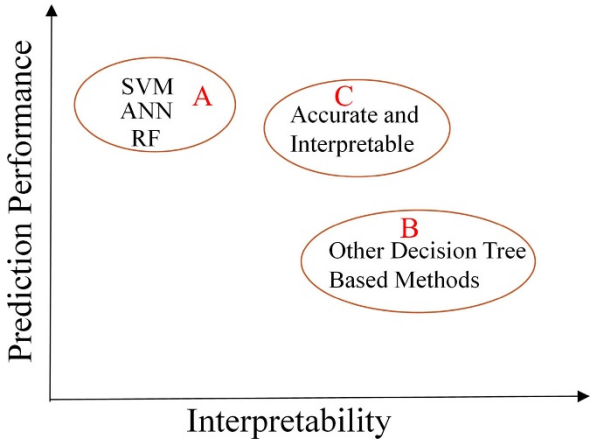

Figure 1 Interpretability versus prediction performance. Algorithms that fall into region C have relatively balanced prediction performance and interpretability. SVM: Support Vector Machines, ANN: Artificial Neural Network, RF: Random Forests.

more interpretable. There is a rich resource of prior work on rule-based learning and feature selection in the fields of bioinformatics and statistical learning. It is beyond the scope of this article to supply a complete survey of the respective areas. Below we review some of the main findings most closely related to this article.

Our contribution

In many biological problems, building a good predictive model that explains the problem well is the ultimate goal of modelling.

High performance and concise representation (i.e., a small rule set and a small feature set) are two important requirements of rule learning methods. Regression tree based methods usually generate a small set of rules. However, their performance is relatively low compared with those using regression with SVM (Support Vector Regression) and random forests. An RF generally has high performance, but generates a large number of rules. It is difficult to interpret the model using a large number of rules. In this article, we take an iterative approach to regularize random forests to obtain refined rules without compromising the performance. RF has an ensemble of regression trees and covers more candidate rules compared with a single decision tree. Regularization keeps only a small number of rules that are the most discriminative. We take an embedded approach with a greedy backward elimination strategy for feature elimination.

We combine rule extraction and feature elimination method iteratively. The result of rule extraction is used for feature elimination. The selected features are then fed into RF and there is 1-norm regularization step to extract important rules. The iterative alternating approach continues until the selected subset of features does not change. Only a few rule learning algorithms are geared toward regression problems as opposed to classification problems. In addition to application to classification case, we apply this iterative approach to another category of learning algorithm - regression rule learning, extending its domain of usage.

It is important to evaluate the quality of the algorithm in terms of prediction performance and interpretability. We use a set of metric to evaluate rule quality as follows:

1 Accuracy: R^2.
2 Variance of accuracy.
3 Interpretability: number of rules.
4 Interpretability: number of variables used in rule.
5 Robustness to noise.

Methods

In this section, we describe the proposed method. First, we present an approach to find the \right" trade off between prediction performance and model complexity using regularization. We then describe our approach by showing a mapping of the forest generated by RF to rule space where many of rules are being removed by 1-norm regularization. Then we present several metrics for evaluation of accuracy and interpretability respectively.

Balancing accuracy and model complexity with regularization

Machine learning algorithms normally learn a function f from input x to output y, that is,

$$y = f(x).$$

A loss function $L(x, y, f)$ is minimized with respect to x, y, and f. The loss function usually takes the form of error penalty, for example, the squared error:

$$L(x, y, f) = (y - f(x))^2$$

which aims at achieving low error rate on training data. It is common that model constructed this way works very well on training data, but not on test data. This is called overfitting. To avoid the overfitting problem, we can add a complexity penalty to the loss function, for example, L_1 regularization:

$$(y - f(x))^2 + \lambda \| w\|_1$$

where w is parameter in the model, λ is tuning parameter to balancing the accuracy and complexity. It can generates relatively less complex model comparing with the previous one. In this article, we use 1-norm regularization. Due to the sparse solution of 1-norm regularization, the model constructed above is much simplified. 1-norm

regularization has been widely applied in statistics and machine learning, e.g., [4,5], and [6]. The above optimization can be solved by a linear program solver (LP).

Rule elimination using 1-norm regularization from random forests mapped rules

From training samples, we can construct a random forest. As the path from a root node to a leaf node in a decision tree is interpreted as a regression rule, a random forests is equivalently represented as a collection of regression rules. Because each sample traverses each tree from root node to one and only one leaf node, we define a feature vector to capture the leaf node structure of a RF. For sample $\mathbf{x}_i$, the corresponding feature vector that encodes the leaf node assignment is defined as $\mathbf{X}^i = [X_1, \cdots, X^q]^T$ where q is the total number of leaf nodes in the forest,

$$X_i = \begin{cases} a_j \text{ if } \mathrm{x}_i \text{ reaches the } j \text{ - th leaf node,} \\ 0 \text{ otherwise.} \end{cases} \quad (1)$$

$$i = 1, \cdots, l, \; j = 1, \cdots, q.$$

where a_j is the target value at leaf node j. We call the space of $\mathbf{X}_i$'s the rule space. Each dimension of the rule space is defined by one regression rule. The above mapping is an extension of binary mapping applied in [7,8] to the regression case.

Using the above mapping, we obtain a new set of training samples in the rule space,

$$\{(\mathrm{X}_1, \gamma_1), (\mathrm{X}_2, \gamma_2), \cdots, (\mathrm{X}_l, \gamma_l)\}.$$

In rule space, we consider the following form

$$\gamma = w^T \mathrm{X} + b. \quad (2)$$

where weight vector $\mathbf{w}$ and scalar b define linear regression function for the sample. The weights in (2) measure the importance of rules: the magnitude of a weight indicates the importance of the rule. Clearly, a rule can be removed safely if its weight is 0. Rule elimination is therefore formulated as a problem of learning the weight vectors.

We use the technique described in previous section, consider the following learning problem using 1-norm regularization:

$$\begin{aligned} &\min_{\mathrm{w},\xi_i} \left(\lambda ||\mathrm{w}||_1 + \sum_{i=1}^{l} \xi_i \right) \\ &\text{s.t. } |\mathrm{w}^T \mathrm{X}_i + b - \gamma_i| \le \xi_i \\ &\qquad \xi_i \ge 0, \; i = 1, \cdots, l. \end{aligned} \quad (3)$$

The solution to the above optimization problem is usually sparse, controlled by regularization parameter λ. λ is chosen by cross validation on the training set. Rules with zero weights w can be removed. Figure 2 illustrate the process shown in this section.

Combined rule extraction and feature elimination

It is assumed that only important features are kept in the remaining rule. Features that do not appear in the rules extracted using (3) are removed because they have no or little effect on the regression problem. In this way, we can select rules and features together.

It is possible to further select rules from a RF built on the selected features to get a more compact set of rules. This motivates an iterative approach. Features selected in the previous iteration are used for constructing a new RF. A new set of rules is then extracted from the new RF. This process continues until the selected features do not change.

Figure 3 illustrate the overall workflow of the algorithm.

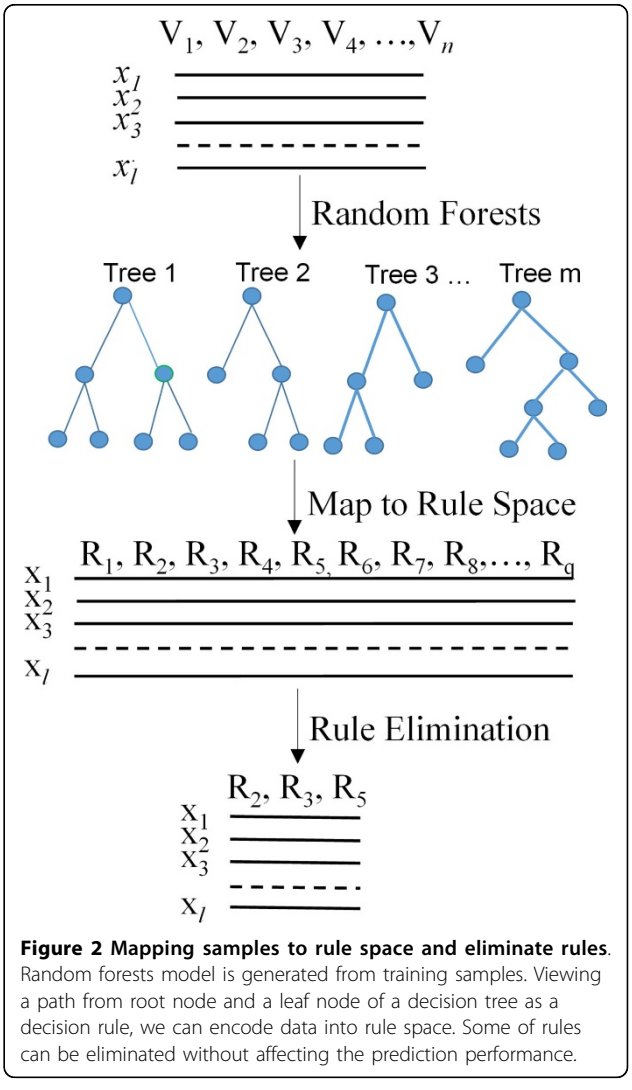

Figure 2 Mapping samples to rule space and eliminate rules. Random forests model is generated from training samples. Viewing a path from root node and a leaf node of a decision tree as a decision rule, we can encode data into rule space. Some of rules can be eliminated without affecting the prediction performance.

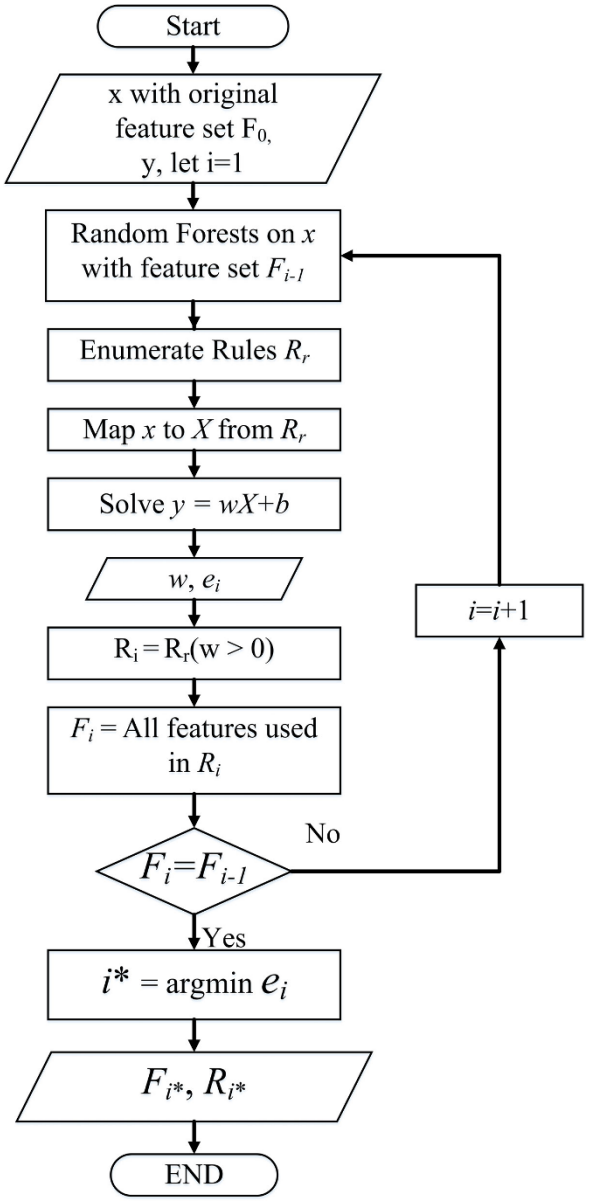

Figure 3 Flowchart of the algorithm. The whole process contains an iterative rule generation and elimination as well as feature elimination until there is no feature eliminated.

Evaluation of results

R squared [9] statistics measures the goodness of fit of the models to the data. It is used to describe how well the predictions fit on test data. Let y' denote prediction values from the algorithm, $\bar{y}$ denote mean value of target variable, the formulation of R squared is:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} \quad (4)$$

where

$SS_{err} = \sum (y_i - y')^2,\ SS_{tot} = \sum (y_i - \bar{y})^2,\ i = 1, \ldots, n,\ n$ is number of test samples. An R squared value closer to one indicates better performance. A simple evaluation on the quality of a regression algorithm is the standard deviation of R squared based on multiple runs. For the interpretability, a small set of rules and concise rules are naturally easier for human to interpret.

In general, random forests classification is more robust against noise compared with many other methods [10]. There is a limited research, however, on whether random forests regression based methods are also robust. One straightforward method is to introduce some noise into the data and then compare the difference between R squared with and without noise. The smaller the difference is, the more robust the algorithm is to noise.

Results and discussion

Datasets

In this section, we first describe the data sets used. We then present detailed results and discussion.

We first test our method on an artificial data set. The data is illustrated in Figure 4. The target values are 1 through 6 corresponding to different shapes and colors. For each target value, 100 samples are generated according to different mean values with standard deviation of 0.5. The relationship between target variable and input variable X1 and X2 is nonlinear.

We then applied our method to several data sets from real applications summarized in Table 1. The first data set is Stockori flowering time data set [11]. The flowering time of 697 plants were collected. The prediction of flowering time is based on 149 genotypes of the plants.

The Parkinson's Telemonitoring data set [12] contains biomedical voice measurement from 42 people with early-stage Parkinson's disease. There are 5875 total voice recordings. The goal is to predict total Uni-fied Parkinson's Disease Rating Scale (UPDRS) scores from the voice measures and other features of patients. Breast Cancer Wisconsin (Prognostic) data set [13] is constructed using a digitized image of a fine needle aspirate (FNA) of a breast mass from breast cancer patients. Characteristic features are computed from the images. The prediction is the recurrence time or disease-free time after treatment. The Relative location of computed tomography (CT) slices on axial axis data set [14] consists of 384 features extracted from CT images. These features are derived from two histograms in polar space. The response variable is relative location of an image on the axial axis ranging from 0 to 180 where 0 denotes the top of the head and 180 the soles of the feet. We randomly choose 2140 CT images for the analysis. The above three data sets are retrieved from University of California, Irvine (UCI) repository [15].

The Seacoast data set is a collection of sensor readings about different biochemical concentrations under various

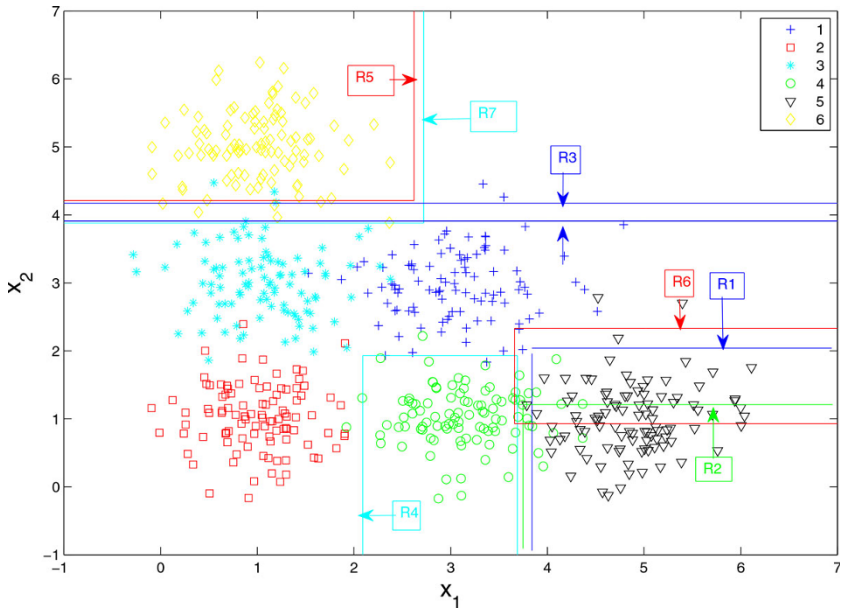

Figure 4 Illustration on artificial data set. Artificial data set described by dots and regression result demonstrated by partition and target prediction in text box.

Table 1 Some statistics of data sets.

Data	Number of Samples	Number of Features
Stockori Floweringtime	697	149
Parkinsons Telemonitoring	5875	19
Breast Cancer Wisconsin (Prognostic)	198	32
Relative location of computed tomography (CT) slices on axial axis	2140	384
Seacoast	2250	16
TCGA Glioblastoma multiforme	427	12042

humidity and temperatures. Concentrations of the biochemical can be inferred from sensor responses using our approach. The data set is pre-processed by normalizing raw sensor responses, calibrating sensor data according to standard no biochemical input conditions and according to the time delay in the sensor response, if available. Humidity levels and temperatures are also factored out first by using a regression based approach. This results in sensor responses being in the same scale. 2250 times points are sampled and used.

The TCGA Glioblastoma multiforme (GBM) data is downloaded from The Cancel Genome Atlas (TCGA) data portal (https://tcga-data.nci.nih.gov/tcga/). 548 gene expression profiles were retrieved from the Broad Institute HT HG-U133A platform (Affymetrix, Santa Clara, CA, USA). Each gene expression profile consists of normalized expression data of 12042 genes. The survival information of patients is retrieved from TCGA clinical data. After removing gene expression samples with unknown survival information, 427 samples were used in our analysis.

Results on artificial data set

Random forests regression generates 11700 rules with R^2 of 0.87. Our method gets 7 rules with R^2 of 0.66. There is not too much loss in the prediction performance. The predicted rules are as follows:

1 **IF** $x_2 \leq 2.04$ and $x_1 > 3.84$ **THEN** $y = 5$
2 **IF** $x_2 \leq 1.21$ and $x_1 > 3.75$ **THEN** $y = 5$
3 **IF** $x_2 \leq 4.17$ and $x_2 > 3.91$ **THEN** $y = 6$
4 **IF** $x_1 \leq 3.68$ and $x_2 \leq 1.93$ and $x_1 > 2.09$ **THEN** $y = 4$
5 **IF** $x_2 > 4.21$ and $x_1 \leq 2.62$ **THEN** $y = 6$
6 **IF** $x_2 \leq 2.33$ and $x_2 > 0.93$ and $x_1 > 3.66$ **THEN** $y = 5$
7 **IF** $x_2 > 3.88$ and $x_1 < 2.72$ **THEN** $y = 6$.

They are also illustrated in Figure 4. Numbers in text boxes are prediction values of target variable. Lines generated from rules partition the original space. Many of these rules align well with the partition. Noted that multiple run

of our approach generates different sets of rules. The number of extracted rules also changes. The partitions in those rules align well with the partition also.

Results on different data sets

The following tables present the result of our proposed methods on different data sets. Results are from test data.

From Table 2, we can see that in all data sets, the number of rules is reduced significantly comparing to random forests yielding less than 1% of the original number of rules in the forest. At the same time, the performance measured by R^2 does not change too much. In most data sets, except Parkinson's Telemonitoring data set, RF gives the best performance. Support vector regression is the least competitive in the cases we tested. Our approach stands somewhere in the middle. Note that on Stockori flowering time data set, the target variable, flowering time, is ordered. Here we simply treat it as numbers. The performance is comparable with RF. In Breast Cancer Wisconsin (Prognostic) data set, the predictive performance is low indicating it is a hard problem. Our approach does not work well on this data set either. It may be resulted from over pruning the rules.

The standard deviation on the R^2, number of rules selected, number of features selected demonstrates that the methods are stable on most of these data sets. The standard deviation of R^2 is obtained from the average of R^2 over ten runs.

Table 2 Results on different data sets.

Numbers after ± are standard deviation. SVR is support vector regression.

	Random Forests	**Our Approach**	**SVR**
			Stockori Flowing Time
R^2	0.54 ± 0.00	0.45 ± 0.05	0.28 ± 0.03
Number of Rules Selected	66020 ± 187	348 ± 33	NA
Number of Features Used in a Rule	8.8 ± 1.9	7.5± 1.74	NA
Number of Features Selected	149 ± 0	135 ± 31	149 ± 0
			Parkinson's Telemonitoring
R^2	0.15 ± 0.02	0.06 ± 0.02	0.17 ± 0.02
Number of Rules Selected	644789 ± 414	3796 ± 0	NA
Number of Features Used in a Rule	9.72± 2.14	7.4 ± 1.86	NA
Number of Features Selected	19 ± 0	19 ± 0	19 ± 0
			Breast Cancer Wisconsin (Prognostic)
R^2	0.04 ± 0.02	-0.19 ± 0.16	-0.04 ± 0.04
Number of Rules Selected	43907 ± 58	126 ± 2	NA
Number of Features Used in a Rule	7 ± 3	3 ± 1.49	NA
Number of Features Selected	32 ± 0	31 ± 1	32 ± 0
			Relative location of CT slices on axial axis
R^2	0.92 ± 0.01	0.77 ± 0.09	0.26 ± 0.00
Number of Rules Selected	172984 ± 143	901 ± 15	NA
Number of Features Used in a Rule	12 ± 3.12	8 ± 2.53	NA
Number of Features Selected	384 ± 0	20 ±± 5	384 ± 0
			Seacoast
R^2	0.64 ± 0.02	0.59 ± 0.10	-0.19 ± 0.00
Number of Rules Selected	120771 ± 161	385 ≤ 5	NA
Number of Features Used in a Rule	14 ±± 3	6 ± 1.91	NA
Number of Features Selected	16 ± 0	16 ± 0	16 ± 0
			TCGA Glioblastoma multiforme
R^2	0.04 ± 0.01	-1.94 ± 0.67	-0.09 ± 0.00
Number of Rules Selected	53539 ± 31344	279 ± 6	NA
Number of Features Used in a Rule	3 ± 2	2 ± 1	NA
Number of Features Selected	12042 ± 0	2 ± 1	12042 ± 0

One example rule set from top five rules based on absolute value of weight in Breast Cancer Wisconsin (Prognostic) data set are as follows:

1 **IF** $v_{22} > 30.27$ and $v_1 \leq 17.23$ and $v_5 > 0.09$ and $v_{11} > 0.24$ and $v_{25} \leq 0.16$ and $v_{14} > 28.38$ and $v_{20} > 0.00$ and $v_{20} \leq 0.01$ **THEN** $y = 64.5$
2 **IF** $v_{12} \leq 1.17$ and $v_9 \leq 0.18$ and $v_3 > 88.13$ and $v_{16} > 0.02$ and $v_{21} > 23.37$ **THEN** $y = 57$
3 **IF** $v_4 > 814.40$ and $v_{12} \leq 0.70$ **THEN** $y = 101.33$
4 **IF** $v_{17} \leq 0.05$ and $v_{30} \leq 0.10$ and $v_{19} \leq 0.01$ and $v_{31} > 0.70$ and $v_{16} \leq 0.03$ and $v_{23} \leq 130.75$ **THEN** $y = 69.25$
5 **IF** $v_{23} > 123.70$ and $v_{29} > 0.26$ and $v_2 \leq 18.43$ and $v_{17} > 0.03$ **THEN** $y = 109.2$.

where v_1 is mean radius, v_2 is mean texture, v_3 is mean perimeter, v_4 is mean area, v_5 is mean smoothness, v_9 is mean symmetry, v_{11} radius standard error (SE), v_{12} is texture SE, v_{14} is area SE, v_{16} is compactness SE, v_{17} is concavity SE, v_{19} is symmetry SE, v_{20} is fractal dimension SE, v_{21} is worst radius, v_{22} is worst texture, v_{23} is worst perimeter, v_{25} is worst smoothness, v_{29} is worst symmetry, v_{30} is worst fractal dimension, and v_{31} is tumor size. Among these rules, size, shape, and texture features occur more often than other features indicating these features are more important than other features in deciding breast cancer. This result is similar to conclusion made in [16] and [17].

To illustrate how prediction values matched true values, we use an approach similar to [18], which was used for clustering analysis. Here we partition the target values in different intervals, and then count how many samples fall into the same interval for both prediction and true values. The resulting confusion matrix can be visualized to get an idea how they match. Figure 5 shows that most of the sample matches are in the diagonal of the matrix which indicate correct match.

Top ten rules are extracted from Seacoast data based on absolute value of weight of rules. Figure 6 shows the number of occurrence of sensors in top ten rules. The sensors are numbered from 1 to 16 accordingly: C0 = 0 MSS556 cp29I Ethyl Cellulose, C1 = 1 MSS556 Ethyl Cellulose, C2 = 2 MSS556 2STH162 (HC), C3 = 3 MSS556 2STH162 (HC), C4 = 4 MSS556 PECH, C5 = 5 MSS556 PECH, C6 = 6 MSS556 PEVA 40%, C7 = 7 MSS556 PEVA 40%, C0 = 0 MSS557 cp27i Ethyl Cellulose, C1 = 1 MSS557 Ethyl Cellulose, C2 = 2 MSS557 2STH162 (HC), C3 = 3 MSS557 2STH162 (HC), C4 = 4 MSS557 PECH, C5 = 5 MSS557 PECH, C6 = 6 MSS557 PEVA 40%, C7 = 7 MSS557 PEVA 40%. From Figure 6, two sensors, C5 = 5 MSS556 PECH and C5 = 5 MSS557 PECH, used more often, suggesting it is more important or effective in determining chemical concentration.

For TCGA Glioblastoma multiforme data set, we can see an interesting result that the number of genes is reduced during iteration, while the number of remaining rules is almost constant after the first iteration. The prediction performance are not good for any of the three algorithms indicating it is a harder problem, and current gene expression profiles may not provide the necessary information for the survival prediction. See Figure 7.

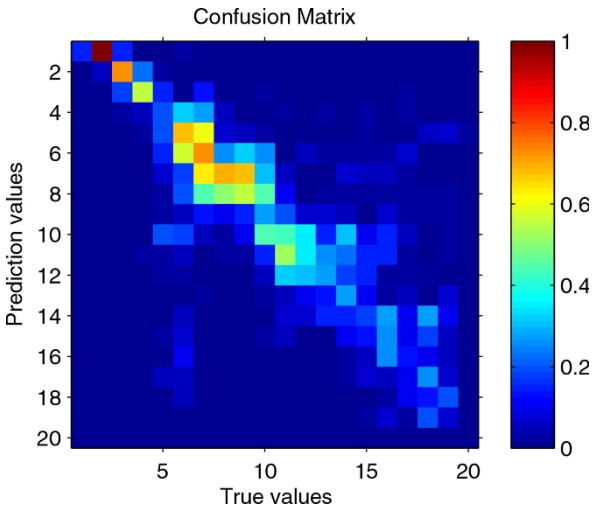

Figure 5 Heatmap of confusion matrix. Target values are partitioned into 20 intervals. True values and prediction values are then discretized into those intervals. Confusion matrix between true values and prediction values is formed by counting the number of matches within each interval. The counts are divided by the maximum count.

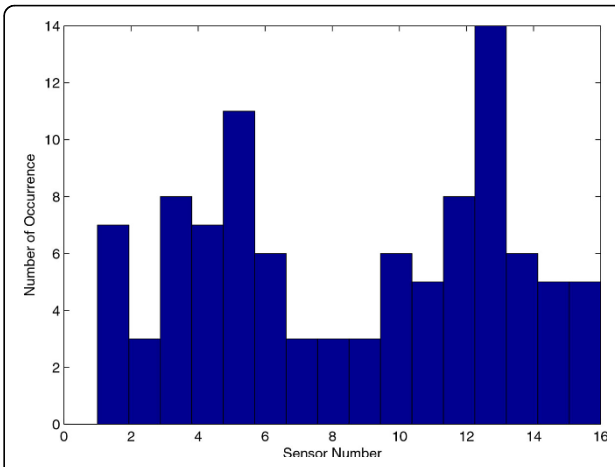

Figure 6 Number of occurrence of sensors in top ten rules. Top ten rules are extracted according to absolute value of weights. The number of occurrence of each sensor is enumerated.

Results with noisy data

To illustrate the robustness of our approach on noisy data, we randomly add some Gaussian noise in the Stockori flowering time data set with probability 0.3 for each

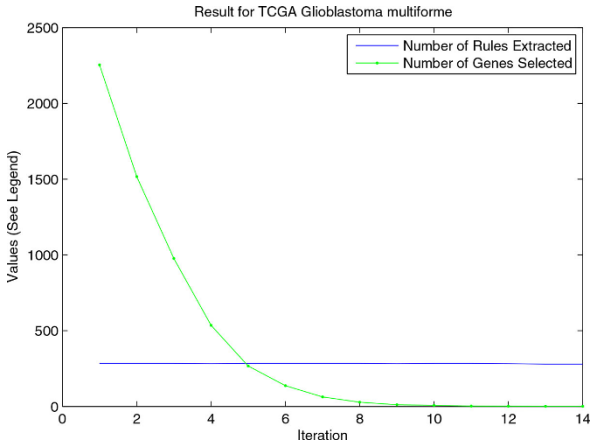

Figure 7 Dynamics of number of rules and genes during iteration. Number of rules extracted and number of genes selected during the iterations of our approach.

feature in a sample. Storckori flowering time was randomly sampled for training and testing sets. The sets were used for the experiments. 10 runs were done for each method. The mean of Gaussian noise is 0.5, while its standard deviation is 0.2. From Table 3, we can see that support vector regression has the highest p value on paired t test on difference in mean R^2s of SVR on data without noise and data with noise, it was not affected too much by Gaussian noise. But its R^2 is still the lowest among all three methods. The p value shows that there is no statistical significance between results with noise and without noise. Our approach has similar values of R^2 compared with those of random forests. Increasing the probability of noise from 0.3 to 1, both random forests and the proposed approach are affected by the increased noise level.

Conclusion

We propose to use an ensemble of decision rules generated from random forests and 1-norm regularization to balance prediction performance and interpretability of regression problems. The method selects a small number of rules (using a small number of features) while retaining performance comparable to RF, better than SVR in most cases.

Due to decision trees' ability handling mixed data type, our approach is able to handles data with mixed type.

Table 3 Result on stockori flowering time data set with noise.

Numbers after ± are standard deviation. SVR is support vector regression.

	Random Forests	Our Approach	SVR
R^2	0.43 ±0.02	0.36 ± 0.07	0.24 ± 0.05
D	0.13	0.1	0.01

We also study robustness of our approach in the presence of noise. The prediction performance is still comparable with random forests in terms of performance within small amount of Gaussian noise.

Regression problems are generally harder than classification problems both in terms of prediction performance and interpretability [8]. Therefore, care should be taken when interpreting the results.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

YC and DW provided guidance and planning for the project. SL produced the program, ran the experiment, and wrote the manuscript. SD, SP, and TM contributed in preparing data and discussions. XD participated in design of experiment. All authors read and approved the final manuscript.

Acknowledgements

This work is supported in part by the US National Science Foundation under award numbers EPS-0903787 and EPS 1006883.

Declarations

The full funding for the publication fee came from Mississippi Experimental Program to Stimulate Competitive Research (EPSCoR).
This article has been published as part of *BMC Systems Biology* Volume 8 Supplement 3, 2014: IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2013): Systems Biology Approaches to Biomedicine. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcsystbiol/supplements/8/S3.

Authors' details

[1]Department of Computer and Information Science, University of Mississippi, Weir Hall 201, 38677, University, MS, U.S.A. [2]Department of Chemistry, Mississippi State University, Box 9573, 37962, Mississippi State, MS, U.S.A. [3]Seacoast Science, Inc, 2151 Las Palmas Dr., Suite C, 92011, Carlsbad, CA, U.S.A. [4]Department of Mathematics, University of Mississippi, Hume Hall 305, 38677, University, MS, U.S.A.

Published: 22 October 2014

References

1. Cortes C, Vapnik V: **Support-Vector Networks.** *Machine Learning* 1995, **20(3)**:273-297[http://dx.doi.org/10.1023/A:1022627411411].
2. McCulloch W, Pitts W: **A logical calculus of the ideas immanent in nervous activity.** *The bulletin of mathematical biophysics* 1943, **5(4)**:115-133.
3. Breiman L: **Random Forests.** *Maching Learning* 2001, **45**:5-32.
4. Tibshirani R: **Regression Shrinkage and Selection Via the Lasso.** *Journal of the Royal Statistical Society, Series B* 1996, **58**:267-288.
5. Zhu J, Rosset S, Hastie T, Tibshirani R: **1-norm Support Vector Machines.** In *Advances in Neural Information Processing Systems 16: December 8-13, 2003; Vancouver and Whistler, British Columbia, Canada.* Thrun S, Saul LK, Schölkopf B 2003, 49-56.
6. Zou H: **An Improved 1-norm SVM for Simultaneous Classification and Variable Selection.** *Journal of Machine Learning Research -Proceedings Track* 2007, **2**:675-681.
7. Liu S, Chen Y, Wilkins D: **Large Margin Classifiers and Random Forests for Integrated Biological Prediction on Mixed Type Data.** *Proceedings of the 7th Annual Biotechnology and Bioinformatics Symposium (BIOT): October 14-15, 2010, Lafayette, Louisiana, USA* 2010, 11-19.
8. Liu S, Patel RY, Daga PR, Liu H, Fu G, Doerksen RJ, Chen Y, Wilkins DE: **Combined rule extraction and feature elimination in supervised classification.** *IEEE Transactions on NanoBioscience* 2012, **11(3)**:228-236.
9. Steel R, Torrie J: *Principles and procedures of statistics, with special reference to the biological sciences.* New York: McGraw-Hill; 1960.
10. Hamza M, Larocque D: **An Empirical Comparison of Ensemble Methods Based on Classification Trees.** *Journal of Statistical Computation and Simulation* 2005, **75**:629-643.

11. Stockori : **Stockori qtl dataset.**[http://agbs.kyb.tuebingen.mpg.de/wikis/bg/stockori.zip].
12. Tsanas A, Little M, McSharry P, Ramig L: **Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests.** *IEEE Transactions on Biomedical Engineering* 2010, **57(4)**:884-893.
13. Street WN, Mangasarian OL, Wolberg WH: **An Inductive Learning Approach to Prognostic Prediction.** In *Proceedings of the Twelfth International Conference on Machine Learning: July 9-12, 1995; Tahoe City, California, USA.*. Burlington: Morgan Kaufmann;Prieditis A, Russell SJ 1995:522-530.
14. Graf F, Kriegel HP, Schubert M, Pölsterl S, Cavallaro A: **2D Image Registration in CT Images Using Radial Image Descriptors.** In *Medical Image Computing and Computer-Assisted Intervention.*. Berlin Heidelberg: Springer-Verlag;Fichtinger G, Martel A, Peters T 2011:607-614, Lecture Notes in Computer Science, vol. 6892.
15. Frank A, Asuncion A: 2010.
16. Wolberg WH, Nick Street W, Mangasarian OL: **Importance of Nuclear Morphology in Breast Cancer Prognosis.** *Clinical Cancer Research* 1999, **5(11)**:3542-3548.
17. Narasimha A, Vasavi B, Harendra Kumar M: **Significance of nuclear morphometry in benign and malignant breast aspirates.** *International Journal of Applied and Basic Medical Research* 2013, **3**:22-26.
18. Monti S, Tamayo P, Mesirov J, Golub T: **Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data.** *Machine Learning* 2003, **52(1-2)**:91-118.

doi:10.1186/1752-0509-8-S3-S5
Cite this article as: Liu *et al.*: Learning accurate and interpretable models based on regularized random forests regression. *BMC Systems Biology* 2014 **8**(Suppl 3):S5.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central