SAND2017-13111C

Exceptional service in the national interest

Sandia National Laboratories

Probability Series Expansion Classifier that is Interpretable by Design

Sapan Agarwal[1], Corey M. Hudson

Overview

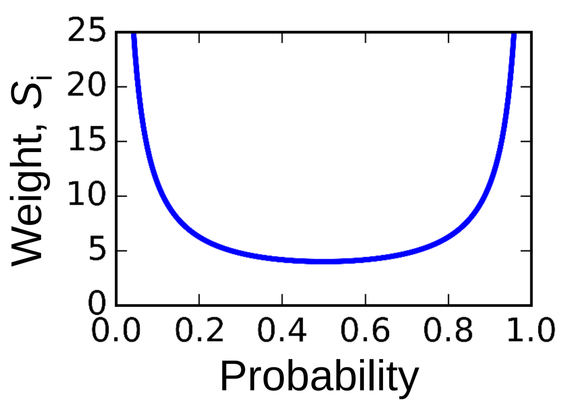
This work presents a new classifier that is specifically designed to be fully interpretable. This technique determines the probability of a class outcome, based directly on probability assignments measured from the training data. The accuracy of the predicted probability can be improved by measuring more probability estimates from the training data to create a series expansion that refines the predicted probability. We use this work to classify four standard datasets and achieve accuracies comparable to that of Random Forests. Because this technique is interpretable by design, it is capable of determining the combinations of features that contribute to a particular classification probability for individual cases as well as the weightings of each of combination of features

Weight P=0 and P=1 More Strongly

Features that predict Y′ more strongly should be weighted more strongly.

Add weight S_i, to the average

$$S_i = \frac{1}{P_i' \times (1 - P_i')} \qquad P' = \frac{\sum_i P_i' \times S_i}{\sum_i S_i}$$



Alternatively tried is logarithmic scaling, but empirically this does not work as well

$$S_i = \frac{\log(P_i'/(1 - P_i'))}{P_i' - 0.5}$$

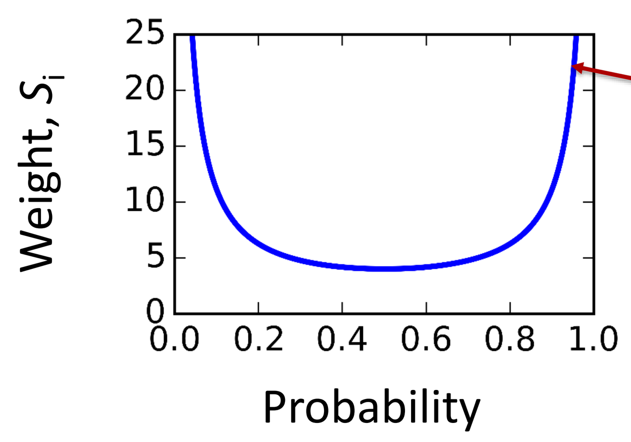
Account for Noise: Regularization

Each probability estimate, P_i, P_{ij}, etc is backed by some number of training examples, n

Limit P=0/1 scaling based on n.

- Limit P_i used to calculate S_i to $\left[\frac{n_{err}}{n}, 1 - \frac{n_{err}}{n}\right]$
  - n_{err} is a meta-parameter around 1, represents average number of examples that might be wrong due to noise
  - If $n < n_{err}$, set $p_i = 0.5$

$$S_i = \frac{1}{P_i' \times (1 - P_i')}$$



Need to avoid large weights due to noise, i.e., P_i=1 based on 1 data point

Additionally add a weight based on the number of samples

Define n_{max} = # of training examples to needed to overcome noise

Scaling factor = $N_i = \min(\sqrt{n_i}, \sqrt{n_{max}})$

- Noise tends to scale with variance, proportional to sqrt (N)

$$P' = \frac{\sum_i P_i' \times S_i \times N_i}{\sum_i S_i \times N_i}$$

If none of the probabilities are backed by significant data, use the parent probability:

$$P' = \frac{\max_i(N_i)}{N} \times \frac{\sum_i P_i' \times S_i \times N_i}{\sum_i S_i \times N_i} + \left(1 - \frac{\max_i(N_i)}{N}\right) \times P$$

Add a Weight Based on Feature Usefulness

Optionally estimate usefulness by how much each feature changes the probability around the decision function

$$U_i = \frac{P_i' - P}{P_i' + P} \qquad P' = \frac{\max_i(N_i)}{N} \times \frac{\sum_i P_i' \times S_i \times N_i \times U_i}{\sum_i S_i \times N_i \times U_i} + \left(1 - \frac{\max_i(N_i)}{N}\right) \times P$$

[1]Contact Information – Sapan Agarwal - sagarwa@sandia.gov

Want a Fast, Adaptable, Explainable Classifier

Find probability of outcome $Y = Y'$ given measured features $x_1 = x_1', x_2 = x_2', x_3 = x_3'$

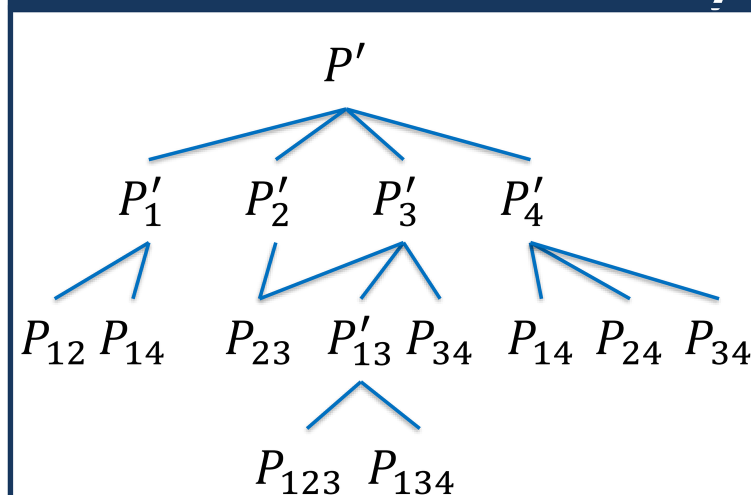
From training data, measure the following probabilities:

$$P(Y = Y') = 0.8 \qquad P(Y = Y' | x_3 = x_3') = 0.7$$
$$P(Y = Y' | x_1 = x_1') = 0.001 \qquad P(Y = Y' | x_1 = x_1', x_2 = x_2') = 0$$
$$P(Y = Y' | x_2 = x_2') = 0.9$$

How do we combine these to get the best overall probability estimate:
$P(Y = Y' | x_1 = x_1', x_2 = x_2', x_3 = x_3')$?

- No Good 1st Principles Method
- Bayesian methods degenerate to values of 0 and 1
- Try averaging Probabilities: $P(Y = 1 | x_1 = x_1', x_2 = x_2', x_3 = x_3') \approx \frac{\sum_i P(Y=1 | x_i = x_i')}{\sum_i 1}$

Hierarchically Average the Probability



$$P' = \frac{\sum_i P_i'}{\sum_i 1} \qquad P = P(Y = Y'),\ P' = \text{estimate of } P$$

$$P_i' = \frac{\sum_{j \neq i} P_{ij}'}{\sum_{j \neq i} 1} \qquad P_i = P(Y = Y' | x_i = x_i')$$

$$P_{ij}' = \frac{\sum_{k \neq j \neq i} P_{ijk}}{\sum_{k \neq j \neq i} 1} \qquad P_{ij} = P(Y = Y' | x_i = x_i', x_j = x_j')$$

$$P_{ijk} = P(Y = Y' | x_i = x_i', x_j = x_j', x_k = x_k')$$

To limit model size when training the classifier, selectively choose which probabilities to keep track of based on which features maximize gini coefficient or minimize entropy

Full Model Accounting for all Factors

$$P' = \frac{\max_i(N_i)}{N} \times \frac{\sum_i P_i' \times S_i \times N_i \times U_i}{\sum_i S_i \times N_i \times U_i} + \left(1 - \frac{\max_i(N_i)}{N}\right) \times P$$

$$P_i' = \frac{\max_j(N_{ij})}{N_i} \times \frac{\sum_{j \neq i} P_{ij}' \times S_{ij} \times N_{ij} \times U_{ij}}{\sum_{j \neq i} S_{ij} \times N_{ij} \times U_{ij}} + \left(1 - \frac{\max_j(N_{ij})}{N_i}\right) \times P_i$$

…

Preliminary Results

Dataset	New Classifier	Random Forests
1984 Congressional Voting Records	96.8%	96.8%
E. Coli Promotor Gene Sequences	95.3%	94.3%
SPECT Heart Data	85.0%	85.0%
Lymphography	86.4%	87.2%

Hyper parameters for both classifiers optimized using 4 fold cross validation

Analyze a Misclassified Result

The classifier predicted with 70% probability that this Member of Congress would be a Republican when they are a Democrat.

Features	Counts Rep.	Counts Dem.	Probability Rep.	Weight	Cumulative Weight
x3=1, x9=1, x15=0	16	0	100%	12.5%	12.5%
x9=1, x15=0	16	0	100%	7.0%	19.4%
x3=1, x8=0, x9=1, x14=0	47	0	100%	4.8%	24.2%
x3=1, x7=0, x9=1, x14=0	47	0	100%	4.7%	28.9%
x2=1, x10=1, x11=0	0	58	0%	4.0%	32.9%
x1=1, x2=1, x10=1	0	38	0%	3.1%	36.0%

Feature #	Probability Republican	Weight	Feature #	Probability Republican	Weight
x9=1	95%	17%	x8=0	86%	4%
x3=1	94%	16%	x7=0	83%	4%
x15=0	94%	14%	x6=0	29%	3%
x2=1	14%	8%	x0=0	78%	3%
x10=1	23%	8%	x13=1	55%	2%
x11=0	21%	6%	x4=1	69%	2%
x14=0	89%	5%	x12=1	52%	2%
x1=1	53%	4%	x5=1	56%	2%

Republican features have higher certainty and therefore higher weight

1984 Congressional Voting Records Dataset

Feature Number	House Bill Description
0	Handicapped infants
1	Water project cost sharing
2	Adoption of the budget resolution
3	Physician fee freeze
4	El Salvador aid
5	Religious groups in schools
6	Anti-satellite test ban
7	Aid to Nicaraguan contras
8	Mx missile
9	Immigration
10	Synfuels corporation cutback
11	Education spending
12	Superfund right to sue
13	Crime
14	Duty free exports
15	Export administration act South Africa

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Exceptional service in the national interest

Sandia National Laboratories

Probability Series Expansion Classifier that is Interpretable by Design

Sapan Agarwal and Corey Hudson

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Want a Fast, Adaptable, Explainable Classifier

Sandia National Laboratories

Find the probability of outcome $Y = 1$
given measured features $x_1 = x_1', x_2 = x_2', x_2 = x_3'$

Estimate $P(Y = 1|\, x_1 = x_1', x_2 = x_2', x_3 = x_3')$

From training data, measure the following probabilities:

$$P(Y = 1) = 0.8$$
$$P(Y = 1|\, x_1 = x_1'\,) = 0.001$$
$$P(Y = 1|\, x_2 = x_2'\,) = 0.9$$
$$P(Y = 1|\, x_3 = x_3'\,) = 0.7$$
$$P(Y = 1|\, x_1 = x_1', x_2 = x_2'\,) = 0$$

How do we combine these to get the best overall probability estimate?

No Good 1st Principles Method

- Bayesian methods degenerate to values of 0 and 1
- All measured values are estimates
- One route to determine the importance of features is by averaging the probabilities

$$P(Y = 1|\, x_1 = x_1', x_2 = x_2', x_3 = x_3') \approx \frac{\sum_i P(Y=1|\, x_i = x_i')}{\sum_i 1}$$

- Need to account for many effects:
  - As a P approaches 0 or 1, it should be given more weight
  - Measured probabilities with less supporting data should be given less weight
  - Measured probabilities that give new information should possibly be given more weight

Use a Hierarchical Weighted Average

Preliminary Results

Dataset	New Classifier	Random Forests
1984 Congressional Voting Records	96.8%	96.8%
E. Coli Promotor Gene Sequences	95.3%	94.3%
SPECT Heart Data	85.0%	85.0%
Lymphography	86.4%	87.2%

Hyper parameters for both classifiers optimized using 4 fold cross validation

Use Model to Analyze a Misclassified Result

Sandia National Laboratories

The classifier predicted with 70% probability that this Member of Congress would be a Republican when they are a Democrat.

Features	Counts Rep.	Counts Dem.	Probability Rep.	Weight	Cumulative Weight
x3=1, x9=1, x15=0	16	0	100%	12.5%	12.5%
x9=1, x15=0	16	0	100%	7.0%	19.4%
x3=1, x8=0, x9=1, x14=0	47	0	100%	4.8%	24.2%
x3=1, x7=0, x9=1, x14=0	47	0	100%	4.7%	28.9%
x2=1, x10=1, x11=0	0	58	0%	4.0%	32.9%
x1=1, x2=1, x10=1	0	38	0%	3.1%	36.0%

Feature #	Probability Republican	Weight	Feature #	Probability Republican	Weight
x9=1	95%	17%	x8=0	86%	4%
x3=1	94%	16%	x7=0	83%	4%
x15=0	94%	14%	x6=0	29%	3%
x2=1	14%	8%	x0=0	78%	3%
x10=1	23%	8%	x13=1	55%	2%
x11=0	21%	6%	x4=1	69%	2%
x14=0	89%	5%	x12=1	52%	2%
x1=1	53%	4%	x5=1	56%	2%

Republican features have higher certainty and therefore higher weight

1984 Congressional Voting Records Dataset

Feature Number	House Bill Description
0	Handicapped infants
1	Water project cost sharing
2	Adoption of the budget resolution
3	Physician fee freeze
4	El Salvador aid
5	Religious groups in schools
6	Anti-satellite test ban
7	Aid to Nicaraguan contras
8	Mx missile
9	Immigration
10	Synfuels corporation cutback
11	Education spending
12	Superfund right to sue
13	Crime
14	Duty free exports
15	Export administration act South Africa