APPLYING IMPROVED RANDOM FOREST EXPLAINABILITY

Student : Dandan Cai

Mentor: Dragutin Petkovic

San Francisco State University

Background

- Machine Learning (ML) is becoming an increasingly critical technology in many areas.
- Complexity and frequent "non-transparency" of ML create significant challenges especially in biomedical and health area.
- Explainability and transparency of ML systems are becoming needed more than ever.

Machine Learning Explainability

- ML explainability can be model based and sample based.
- ML explainability is also essential component of powerful "user in the loop" model.
- ML systems that are explainable may achieve many benefits.
  - Increase user trust
  - Improvement in controlling, quality control and maintenance
  - Legal transparency
  - Offer new insights into the analyzed domain

Principles Of AI Development And Deployment Which All Include Transparency And Explainability Are Emerging

- New **EU General Data Protection** laws (GDPR) effective May 2018 – includes strong data privacy and "right to know" how algorithms work (recital 71)
  - https://www.privacy-regulation.eu/en/r71.htm
- Asilomar **23 AI Principles** adopted by CA legislature
  - https://futureoflife.org/ai-principles/
- G20 AI Principles
  - https://www.oecd.org/going-digital/ai/principles/

Research Community And Universities Take Notice Of AI Issues

- Workshops
  - D. Petkovic (Chair), L. Kobzik, C. Re: "Workshop on Machine learning and deep analytics for biocomputing: call for better explainability", Pacific Symposium on Biocomputing PSB 2018, Hawaii
  - Workshop on AI Ethics and Values in Biomedicine - Technical Challenges and Solutions
  - Pacific Symposium on Biocomputing, Hawaii January 3-7, 2020; *Prof. D. Petkovic, Prof. L. Kobzik, Dr. R. Ghanadan*
- DARPA Program for Explainable AI https://www.darpa.mil/program/explainable-artificial-intelligence
- Stanford Human-Centered AI http://hai.stanford.edu/
- New MIT Initiative in AI Ethics http://news.mit.edu/2018/mit-reshapes-itself-stephen-schwarzman-college-of-computing-1015
- San Francisco State Univ. – Certificate in Ethical AI http://cob.sfsu.edu/management/certificate-programs/ai-ethics

Motivation And Goals

- Motivation
  - Our research is to focus on both model and sample explainability of Random Forest (RF) classifiers.
- Goals
  - To improve RF model explainer and develop new RF sample explainers that are designed from the ground with non-ML experts in mind
  - Use User Centered Approach with ML non-experts as main users: simplicity and familiarity. and provide one page tabular output and measures familiar to most users.
  - To test the new explainers on several databases
  - To document and publish

About RF

- What is RF?
  - Random Forest, a popular and powerful ensemble supervised classification method
  - RF training includes built in approximate 3 fold cross validation
  - RF consists of a set of Ntree decision trees voting for the winning class.
  - *mtry* randomly selected features (with replacement) tried at each tree node
  - Ntrees vote for winning class. Cutoff is used to select sensitivity
  - *MDA – mean decrease in accuracy (feature value perturbation measure) measure used to rank the features*
  - Ref: Breiman L, "Random forests," Machine Learning, vol. 45, no. 1, pp.5–32, 2001

Related Work On RF Explainability

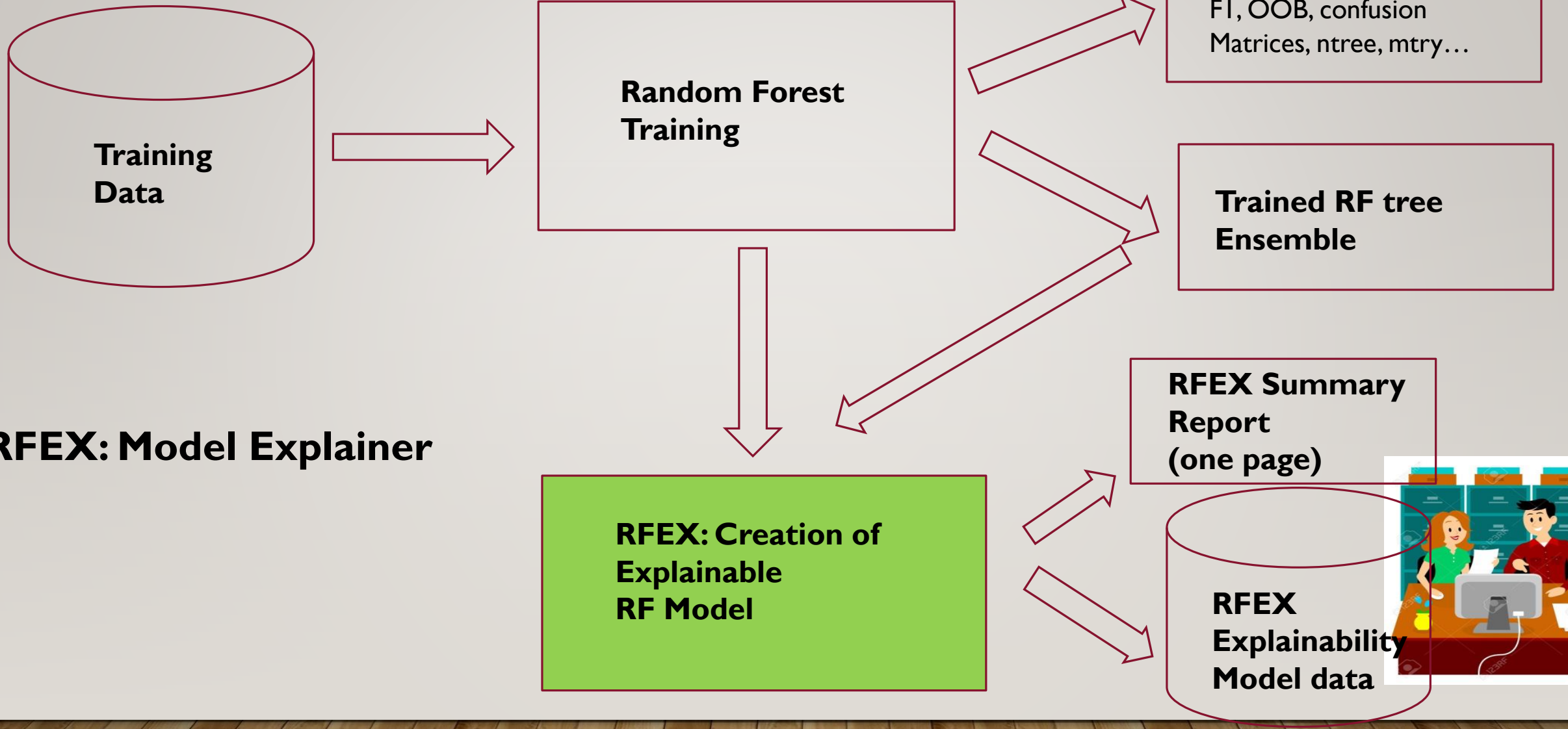
Ref: PACIFIC SYMPOSIUM ON BIOCOMPUTING, HAWAII JANUARY 3-7, 2020; *PROF. D. PETKOVIC, PROF. L. KOBZIK, DR. R. GHANADAN*

- Some work relates to extracting rules from trained Ntrees, but number of rules is too high to be easy to use
- Most use general MDA and not class specific MDA
- Original RFEX approach published by Prof. D. Petkovic at PSB 2018

RF Explainer Approach

- What is RF Explainer?
  - Use training database and standard RF tools/algorithms to produce *base* RF accuracy estimates.
  - RFEX Model explainer: Process trained RF to extract model representation in a summary table – how RF works on totality of data
  - Implemented a *user-centered-approach.* ➔ *users are non-ML experts but domain experts*
    - *Similarity and familiarity* with *report formats* users are used to ➔ One page RFEX explainability summary report in tabular format.

Traditional RF Classification



RFEX: Model Explainer

Sample Explainer Approach

- What is Sample Explainer?
  - RFEX Sample Explainability Data
  - RFEX Sample Summary Table
  - Is designed with the target users in mind.
  - RFEX Sample explainer: for a specific sample, produce RFEX SAMPLE explainer summary table – how RF classifies specific sample

Traditional RF Classification

Training Data

Random Forest Training

Accuracy:
F1, OOB, confusion Matrices, ntree, mtry…

Trained RF tree Ensemble

RFEX Sample Explainer

Sample under investigation

Creation of Sample Explainer Model

Sample Explainer Summary Report (one page)

Sample Explainability Sample Data

Previous Work at SFSU

- Developed RF Explainability Enhancement pipeline (RFEX) as RF Model explainer
  - Petkovic D, Altman R, Wong M, Vigil A.: "Improving the explainability of Random Forest classifier - user centered approach". Pacific Symposium on Biocomputing, Hawaii, 2018
- Applied it to Stanford FEATURE data
- Performed usability experiment with 13 users and demonstrated that RFEX increased their understanding of RF classification

Our Contributions

- Improved original RFEX Model explainer
  - Proposed replacement of p-value measure with Cohen distance
  - Improved RFEX Model summary table
  - Preformed experiments on several DB to verify
- Developed, with Prof. Petkovic, NEW RFEX Sample Explainer (algorithms and tabular representation)
- Developed, several measures to detect "problematic" features"
  - Tested in on several DB
  - Performed sensitivity analysis
- Published SFSU TR and will have poster at PSB 2020

Development Of RFEX Model Explainer

- a) significant improvement in RFEX Model explainer.
  - Choose Cohen Distance over P-value.
  - Add AV/SD [Min/Max] class specific value ranges.
  - Cliques of N features.
  - Tested RFEX Model Summary on several DBs
    - Synthetic data from Sabiha Barlaskar
      - Barlaskar S, Petkovic D: "Applying Improved Random Forest Explainability (RFEX 2.0) on synthetic data", SFSU TR 18.01
    - FEATURE DB from Stanford
    - BC Database UC Irvine

Man Whitney Wilcox (MWW) Test

- Man-Whitney-Wilcox Test is a test of a hypothesis of the distribution of data. It is a non-parametric test of assuming a null hypothesis of class specific samples not being independent from each other. Used in RFEX to measure independence (class separation) of feature values.

- To test the independence, we use the Wilcox.Test function in R to calculate the p-values. If p-value < 0.05, we can reject the null and can say that the feature values are independent of each other.

    - wilcox.test(pos_class_process[,i,], (neg_class_process[,i]))

Cohen Distance

- To determine separation of feature values for + and – class for each feature
- Cohen Distance between feature values of two populations e.g. of + and – class is:
  - Cohen Distance = ABS(AV+ - AV-)/SD

Why Cohen Distance

- We discovered p value test does not work well for large data set (e.g. over 1000 points) and always shows independence
- We performed experiments on synthetic data to prove it
- We confirmed this by literature search e.g. Ref: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444174/
- As per the paper above , we chose Cohen Distance to determine separation of feature values for + and – class for each feature
  - Cohen Distance between feature values of two populations e.g. of + and – class is:
    - Cohen Distance = ABS(AV+ - AV-)/SD
    - Provides better measure which works for large data sets and also gives the level of separation

RFEX Model Explainer: Feature Data (Stanford)

Feature Rank	Feature Name	MDA+ Value	F1 Score Using Top K Features	AV/SD for site class	AV/SD for nonsite class	Cohen's d	Best Pair(F1 score)
1	Solvent_Accessibility_s5	21.9611	N/A	291.98/ 112.35	923.04/ 435.37	1.45	Residue_name_Is_Gly_s 2 (0.8739)
2	Secondary_Structure1_Is _Turn_s3	21.8895	0.8	8.32/2. 02	1.78/2. 71	2.41	Residue_name_Is_Gly_s 2 (0.8617)
3	Residue_Name_Is_Thr_s4	21.6433	0.928	5.66/1. 67	0.85/1. 56	2.88	Residue_name_is_Gly_s 2 (0.8764)
4	Residue_Name_Is_Gly_s3	20.91	0.965	3.44/1. 15	0.49/1. 11	2.57	Residue_Class1_ Is_Unknown_s2 (0.8371)
5	Solvent_Accessibility_s4	20.6998	0.977	167.80/ 94.15	673.39/ 361.95	1.40	Residue_name_Is_Gly_s 2(0.8653)
6	Secodary_Structure_1_Is _Strand_s5	20.6161	0.985	15.25/2 .01	4.29/6. 29	1.74	Residue_name_is_Gly_s 2((0.8679)
7	Residue_Class1_Is_Unkno wn_s3	20.5889	0.986	3.45/1. 16	0.53/1. 20	2.43	Residue_name_is_Gly_s 2(0.8353)
8	Residue_Name_Is_Gly_s2	20.4037	0.990	3.70/1. 51	0.17/0. 60	2.34	Neg_Charge_s2(0.8919)

RFEX Model Explainer: Breast Cancer Data (UC Irvine ML DB)

Feature Rank	Feature Name	MDA value	Cumulative F1 score	AV/SD pos class [MIN, MAX]	AV/SD neg class [MIN, MAX]	Cohen Distance	Top 10 cliques of 2 features
1	Bare Nuclei	17.26	N/A	7.63/3.11 [1, 10]	1.31/1.18 [1, 10]	2.03	[Uniformity of Cell Size, Bare Nuclei] [Bare Nuclei, Normal Nucleoli] [Uniformity of Cell Shape, Bare Nuclei] [Uniformity of Cell Shape, Bland Chromatin] [Uniformity of Cell Size, Single Epithelial Cell Size] [Single Epithelial Cell Size, Bare Nuclei] [Uniformity of Cell Size, Bland Chromatin] [Uniformity of Cell Shape, Normal Nucleoli] [Clump Thickness, Bare Nuclei] [Clump Thickness, Uniformity of Cell Shape]
2	Clump Thickness	13.36	0.904	7.2/2.43 [1, 10]	2.96/1.67 [1,8]	1.74	
3	Uniformity of Cell Shape	13.09	0.932	6.56/2.56 [1, 10]	1.44/0.99 [[1, 8]	2	
4	Bland Chromatin	12.88	0.941	5.98/2.27 [1, 10]	2.10/1.08 [1, 7]	1.71	
5	Uniformity of Cell Size	11.73	0.949	6.57/2.72 [1, 10]	1.33/0.91 [1, 9]	1.93	
6	Marginal Adhesion	9.06	0.951	5.55/3.21 [1, 10]	1.36/0.99 [1, 10]	1.31	
7	Normal Nucleoli	7.89	0.945	5.86/3.35 [1, 10]	1.29/1.06 [1, 9]	1.36	
8	Mitoses	6.80	0.957	2.59/2.56 [1, 10]	1.06/0.50 [1, 8]	0.59	
9	Single Epithelial Cell Size	6.22	0.961	5.30/2.5 [1, 10]	2.12/0.92 [1, 10]	1.27	

RFEX Model for Breast Cancer Discussion

- RF can perform very good classification using all 9 features.
- RFEX Model explainer shows that using only top 3 ranked features one gets over 95% of accuracy using all 9 features (Similar drastic reduction is shown in several other RFEX case studies)

Design And Development Of RFEX Sample Explainer

- Goals
  - Present easy to understand tabular summary of how specific sample has been classified by RF
  - Make it consistent with RFEX Model Explainer
  - Develop rules to detect features that may cause sample to be problematic (outlier) or wrong classification
  - Test on several DB

RFEX Sample Explainer

- RFEX Sample Explainability Data consisting of several global sample level types of information:
  - a) CORRECT_CLASS: Correct (ground truth) class label of tested sample known or assumed by the domain expert
  - b) RF_CLASS_LABEL: Sample class label determined by RF
  - c) VOTE_FRACTION: Fraction of RF trees (relative to ntree of trained classifier) voting for correct class

RFEX Sample Explainer

- RFEX Sample Summary Table shows in one page summary how particular features (topK features from RFEX Model Summary) contribute to RF decisions on tested sample.
  - Feature Rank (from RFEX Model Explainer)
  - Feature Name (from RFEX Model Explainer)
  - Feature MDA rank (from RFEX Model Explainer)
  - Feature values AV/SD for positive class, [Min,Max]
  - Feature values AV/SD for negative class, [Min,Max]
  - Feature Value of Tested Sample
  - Sample Cohen Distance to positive class
  - Sample Cohen Distance to negative class
  - K Nearest Neighbor (KNN) ratio for positive class
  - K Nearest Neighbor (KNN) ratio for negative class

RFEX Sample Explainer Case Study Example: Breast Cancer Data

- Breast cancer databases was obtained from the University of Wisconsin Hospitals
  - https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29
- Number of Instances: 699 (as of 15 July 1992)
- Missing attribute values: 16
  - na.roughfix is used to impute missing values by the random forest model.

Breast Cancer Data Attributes

Attribute	Domain
Clump Thickness	1-10
Uniformity of Cell Size	1-10
Uniformity of Cell Shape	1-10
Marginal Adhesion	1-10
Single Epithelial Cell Size	1-10
Bare Nuclei	1-10
Bland Chromatin	1-10
Normal Nucleoli	1-10
Mitoses	1-10
Class	B(for Benign), M(for malignant)

RFEX Sample Explainer Case Study Example: Breast Cancer Data - GOALS

- Test RFEX Sample Explainer (and verify RFEX Model explainer in doing that)
- Perform analysis of "problematic features" e.g. what features may cause samples to be "marginal or outliers" e.g. those who get low RF tree votes
- Attempt to derive simple rules to identify "problematic features"
- Perturb values of problematic features to average for correct class to see if "correction" would increase RF tree votes
- Above analysis is useful in a typical case of QA and editing of training databases for outliers and for increasing user trust

CASE STUDY
Breast Cancer Sample Explainability-Good Vote

Feature Rank	Feature Name	Feature MDA+ Rankings	Feature AV/SD for Positive Class; [Min, Max]	Feature AV/SD for Negative Class; [Min, Max]	Feature Value of Tested Sample	Sample Cohen Distance to Positive Class	Sample Cohen Distance to Negative Class	K Nearest Positive Neighbor Ratio	K Nearest Negative Neighbor Ratio
1	Clump Thickness	15.10	7.2/2.43 [1, 10]	2.96/1.67 [1,8]	8	0.33	3.02	44/48	4/48
2	Bare Nuclei	14.20	7.63/3.11 [1, 10]	1.31/1.18 [1, 10]	10	0.76	7.36	48/48	0/48
3	Uniformity of Cell Shape	11.11	6.56/2.56 [1, 10]	1.44/0.99 [[1, 8]	7	0.61	5.62	46/48	2/48
4	Bland Chromatin	10.31	5.98/2.27 [1, 10]	2.10/1.08 [1, 7]	7	0.45	4.54	48/48	0/48
5	Uniformity of Cell Size	8.92	6.57/2.72 [1, 10]	1.33/0.91 [1, 9]	8	0.53	7.33	48/48	0/48
6	Marginal Adhesion	7.82	5.55/3.21 [1, 10]	1.36/0.99 [1, 10]	4	0.48	2.67	43/48	5/48
7	Normal Nucleoli	4.83	5.86/3.35 [1, 10]	1.29/1.06 [1, 9]	8	0.64	6.33	48/48	0/48
8	Single Epithelial Cell Size	3.52	5.30/2.5 [1, 10]	2.12/0.92 [1, 10]	10	1.88	8.57	47/48	1/48
9	Mitoses	3.16	2.59/2.56 [1, 10]	1.06/0.5 [1, 8]	7	1.72	11.88	45/48	3/48

CASE STUDY
Breast Cancer Sample Explainability-bad Vote

Feature Rank	Feature Name	Feature MDA+ Rankings	Feature AV/SD for Positive Class; [Min, Max]	Feature AV/SD for Negative Class; [Min, Max]	Feature Value of Tested Sample	Sample Cohen Distance to Positive Class	Sample Cohen Distance to Negative Class	K Nearest Positive Neighbor Ratio	K Nearest Negative Neighbor Ratio
1	Clump Thickness	15.10	7.2/2.43 [1, 10]	2.96/1.67 [1,8]	4	1.32	0.62	37/48	11/48
2	Bare Nuclei	14.20	7.63/3.11 [1, 10]	1.31/1.18 [1, 10]	5	0.85	3.13	36/48	12/48
3	Uniformity of Cell Shape***	11.11	6.56/2.56 [1, 10]	1.44/0.99 [[1, 8]	1	2.17	0.44	1/48	47/48
4	Bland Chromatin ***	10.31	5.98/2.27 [1, 10]	2.10/1.08 [1, 7]	2	1.75	0.09	6/48	42/48
5	Uniformity of Cell Size***	8.92	6.57/2.72 [1, 10]	1.33/0.91 [1, 9]	1	2.05	0.36	3/48	45/48
6	Marginal Adhesion	7.82	5.55/3.21 [1, 10]	1.36/0.99 [1, 10]	3	0.79	1.66	27/48	21/48
7	Normal Nucleoli	4.83	5.86/3.35 [1, 10]	1.29/1.06 [1, 9]	1	1.45	0.27	41/48	7/48
8	Single Epithelial Cell Size***	3.52	5.30/2.5 [1, 10]	2.12/0.92 [1, 10]	1	1.72	1.22	0/48	48/48
9	Mitoses	3.16	2.59/2.56 [1, 10]	1.06/0.5 [1, 8]	1	0.62	0.12	48/48	0/48

RFEX Sample Explainer: Empirical Rule For Detecting Possibly Problematic Features

- Feature is *problematic* IFF
  - Cohen Distance for correct class > Cohen Distance for incorrect class

AND

  - KNN for correct class < 0.5

Sample Explainability
Breast Cancer

- The following experiment takes the top 3 problematic features from lowest vote positive sample and modify its value close to its average value of correct class, modify one feature value at a time and recording the change in the vote.

Problematic Feature	Original Value	Modified Value(average value of its correct class)	Original Vote	Updated Vote
Uniformity of Cell Shape	1	6.56	(0.47, 0.53)	(0.375, 0.625)
Bland Chromatin	2	5.98	(0.47, 0.53)	(0.415, 0.585)
Uniformity of Cell Size	1	6.57	(0.47, 0.53)	(0.3, 0.7)

Sample Explainability
Breast Cancer-- example of perturbing highly ranked feature to incorrect value

- The following experiment takes the highest rank feature that is not problematic for a sample classified correctly as positive but with lowest RF vote, and changes (corrupts) its value to negative class average value and record the vote.

Feature Rank	Original Value	Average Value of Wrong Class	Original Vote	Updated Vote
Clump Thickness	4	2.96	(0.47, 0.53)	(0.505, 0.495)

- Originally the RF model votes 53% for positive class, after modifying the value, RF model votes only 49.5% for positive class, changing this sample from positive to negative.

Sample Explainability Discussion

- Number of features that are identified as problematic correlates with lower RF tree vote of tested sample
- Sensitivity analysis (e.g. perturbation of problematic feature value to average of correct class) shows significant increase of RF tree vote, confirming the influence of specific feature for low RF tree votes
- Errors in highly ranked features (e.gt. Errors in data capture) may cause wrong classification

Sample Explainability Discussion

- RFEX Sample explainer identified problematic features casing low RF votes
- Using simple rule on measures in RFEX Sample explainer, one can identify problematic features from RFEX sample explainer. The more problematic features the sample has, the lower the RF vote count it gets for the correct class
- Problematic feature identification is important to filter out outlier samples, noisy samples or samples with noisy features' or wrong grounds truth classification, which then can ensure better RF training

Tools Used

- R for writing the code and perform random forest engine training
- Sublime for data management
- Excel for data perturbing

Conclusions

- What has been done?
- Improved original RFEX Model explainer
  - Proposed replacement of p-value measure with Cohen distance
  - Improved RFEX Model summary table
  - Preformed experiments on several DB to verify
- Developed, with Prof. Petkovic, NEW RFEX Sample Explainer (algorithms and tabular representation)
- Developed, several measures to detect "problematic" features"
  - Tested in on several DB
  - Performed sensitivity analysis
- Published SFSU TR and will have poster at PSB 2020

Conclusion

- Future work
  - Publish as a poster at PSB 2020 (in progress)
  - Publish research paper in 2020

Also:

  - Test with real users to get their feedback
  - Make Jupyter toolkit for new model and sample explainers

Acknowledgement

- Special thanks to Prof. Petkovic
- Prof. L. Kobzik, Dr. L. Buturovic and Prof. K. Okada for valuable feedback
- Mike Wong's clear explanation of Stanford FEATURE data
- Sabiha Barlaskar 's help at the beginning of the project.
- 13 anonymous usability reviewers.
- The work has partially been supported by NIH grant R01 LM005652 and by SFSU Center for Computing for Life Sciences.