A Novel Meta Learning Framework for Feature Selection using Data Synthesis and Fuzzy Similarity

IEEE World Congress on Computational Intelligence (WCCI) 2020

Zixiao Shen[123], Xin Chen[13] & Jonathan M. Garibaldi[123]

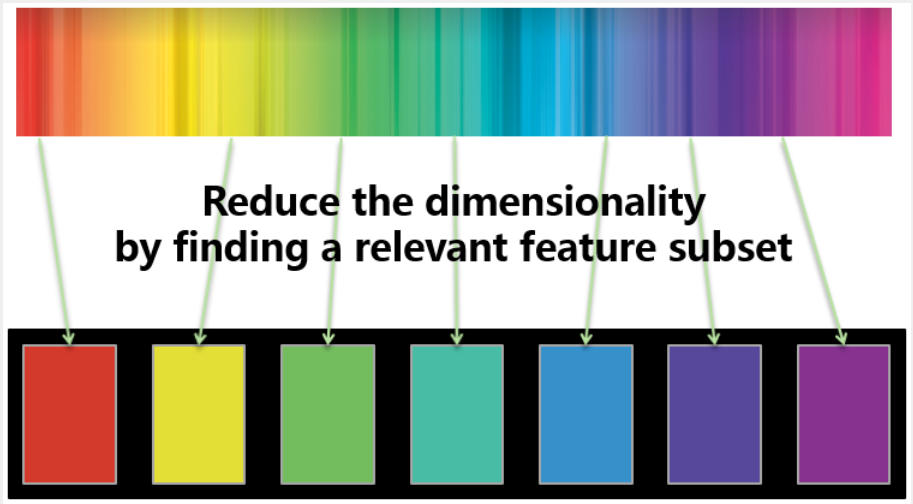
[1] Intelligent Modelling and Analysis Group, School of Computer Science
[2] Lab for Uncertainty in Data and Decision Making (LUCID)
[3] University of Nottingham, Nottingham, NG8 1BB, United Kingdom

Presented by:
Zixiao Shen
22nd July 2020

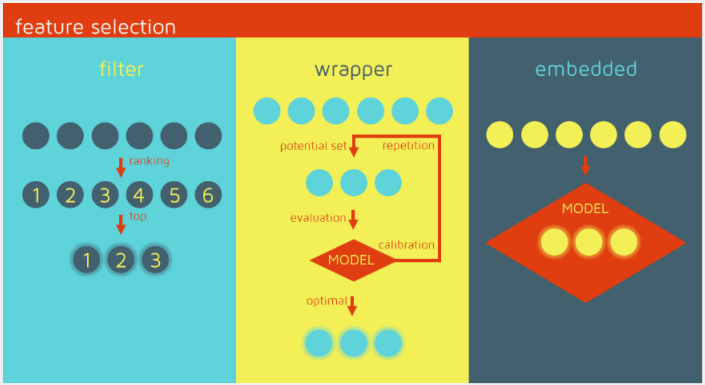
Presentation Outline

What is Feature Selection (FS) ?

Feature selection (FS) is the process of selecting of relevant features for use in model construction.



There exist plenty of different feature selection (FS) methods



- The performance of the various FS methods is data dependent;
- It's not possible to state categorically the optimal FS method for all kinds of data.

Possible Solutions

1. Combination methods using the diversity kinds of FS algorithms[1];
2. Use meta-learning method to choose the best algorithm for a given dataset.

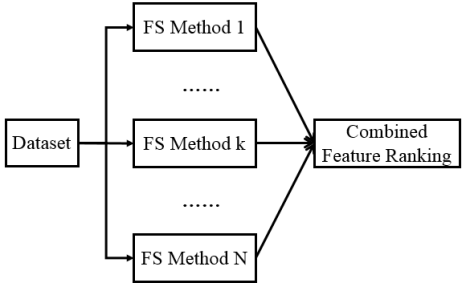


Figure 1: Combined approach for FS

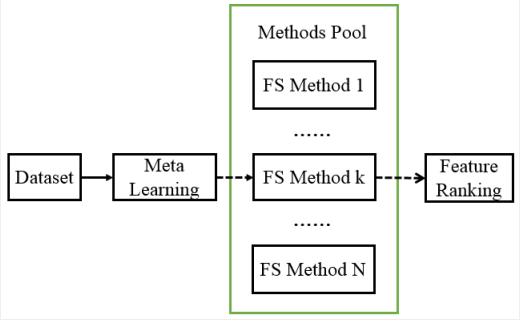


Figure 2: Meta Learning approach for FS

[1] Zixiao Shen, Xin Chen, and Jonathan M Garibaldi. "A Novel Weighted Combination Method for Feature Selection using Fuzzy Sets". In: *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE. 2019, pp. 1–6.

Presentation Outline

Definition

- Meta learning is defined as a process of learning the meta-knowledge to improve model learning using machine learning and data mining methods;
- Two main aspects of research: algorithm selection and parameter selection.

Main Issues of Meta Learning

1. Construction of A Data Repository for Training:
   - Synthesized datasets are proposed to construct a large data repository.
2. Selection of Meta Features:
   - A number of widely used meta features are implemented in our study.
3. Choice of A Recommendation Method:
   - A fuzzy similarity based framework is implemented to achieve the decision making[2].

[2] Zixiao Shen, Xin Chen, and Jon Garibaldi. "Performance Optimization of a Fuzzy Entropy Based Feature Selection and Classification Framework". In: *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. 2018, pp. 1361–1367.

Presentation Outline

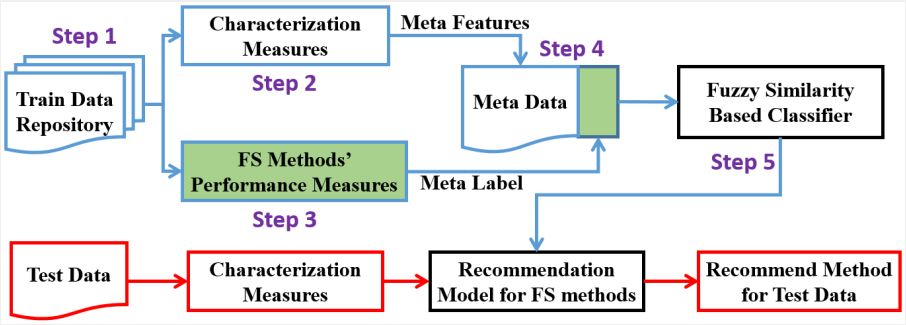


Figure 3: Overall framework of the proposed architecture

1. Generation of a data repository for training;
2. Meta features extraction;
3. FS methods' performance measures;
4. Meta data construction;
5. Recommendation using fuzzy similarity measure.

- Construct a data repository that covers a variety of characteristics using data synthesis;
- Madelon datasets are used on account of its high flexibility and variability;

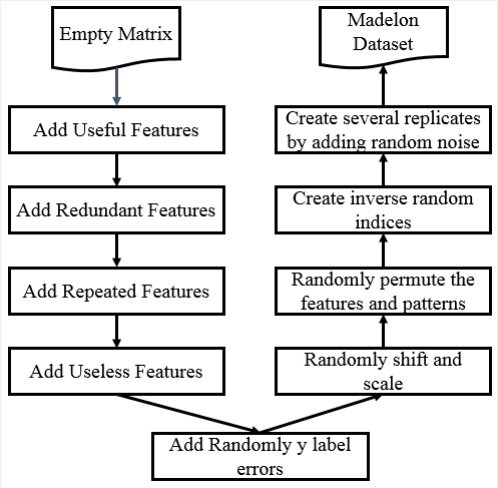


Figure 4: Generation Process of Madelon Dataset

Generation of a Data Repository for Training

- Different kinds of madelon datasets are generated by varying 11 different parameters.

Table 1: Parameters for data synthesis using Madelon dataset

Alias	Meaning	Value Range
P1	Number of Classes	2
P2	Number of Useful Features *(initially drawn to explain the concept)*	[4, 5,..., 20]
P3	Number of Redundant Features *(linearly dependent upon the useful features)*	[0, 1,..., 20]
P4	Number of Repeated Features *(repeating P2 and P3 at random)*	[0, 1,..., 20]
P5	Number of Useless Features *(Drawn at random regardless of class label)*	[0, 1,..., 20]
P6	Number of Samples per Cluster	[10, 11,..., 70]
P7	Number of Cluster per Class	[2, 3,..., 7]
P8	Random Seed	[1, 2,..., 1000]
P9	Factor multiplying the hypercube dimension	[2, 3,..., 10]
P10	Fraction of y labels to be randomly exchanged	[0.01, 0.02, ..., 0.1]
P11	Flag to enable or disable random permutations	[0, 1]

Meta Feature Extraction

- To learn meta features from the synthetic dataset, we extract a set of meta features from M different datasets $D_i, i = 1, ..., M$, each with the number of S_i data samples $(E_1, E_2, ..., E_{S_i})$ and N_i features $(F_1, F_2, ..., F_{N_i})$. The label information is represented using class C $(c_1, c_2, ..., c_{S_i})$ for different data samples.

1. Number of Samples (NS)
2. Number of Features (NF)
3. Avg. Asymmetry of Features (AAF)

$$\frac{3}{N_i}\sum_{j=1}^{N_i}\frac{Mean(F_j) - Med.(F_j)}{Std(F_j)} \quad (1)$$

4. Avg. Correlation of Features (ACF)

$$\frac{2}{N_i(N_i-1)}\sum_{j=1}^{N_i-1}\sum_{k=j+1}^{N_i} Pearson(F_j, F_k) \quad (3)$$

5. Avg. Coef. of Var. of Features (ACVF)

$$\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{Std(F_j)}{Mean(F_j)} \quad (2)$$

6. Avg. Entropy of Features (AEF)

$$\frac{1}{N_i}\sum_{k=1}^{S_i} Entropy(F_j) \quad (4)$$

The best FS method for a given synthetic dataset has been generated as the label.

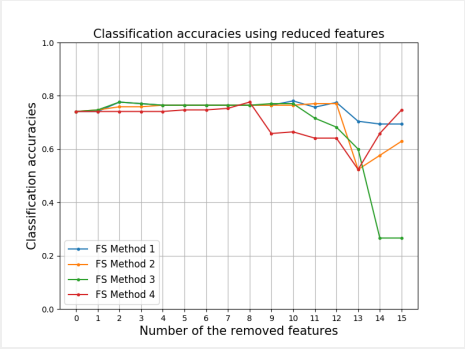


Figure 5: Demonstration of classification accuracies using reduced features

1. Divide the data using 10-fold CV;
2. Implement the candidate FS method to rank the features using the training set;
3. Model the classifier using the training set and make the prediction on the test set;
4. Calculate mean classification accuracy across different folds.

Weighted sum (WS) of the accuracies:

$$WS = \sum Acc. * \%RemovedFeatures$$

- The FS method with the highest WS value is selected as the meta label for the dataset;
- The meta data is constructed by combining the six different meta features MF_p, $(1 \leq p \leq 6)$ and the corresponding meta label for each dataset D_i.

Table 2: Demonstration of Meta Data

Data	**Meta Features**						**Meta Target**
	MF_1	MF_2	...	MF_p	...	MF_6	
D_1	$w_{1,1}$	$w_{1,2}$	...	$w_{1,p}$	...	$w_{1,6}$	Opt_1
D_2	$w_{2,1}$	$w_{2,1}$	...	$w_{2,p}$	...	$w_{2,6}$	Opt_2
...	...	...	...	...	...	...	...
D_i	$w_{i,1}$	$w_{i,1}$	...	$w_{i,p}$	...	$w_{i,6}$	Opt_i
...	...	...	...	...	...	...	...
D_M	$w_{M,1}$	$w_{M,1}$	...	$w_{M,p}$	...	$w_{M,6}$	Opt_M

- Blue line shows data flow for training process;
- Red line shows data flow for testing process.

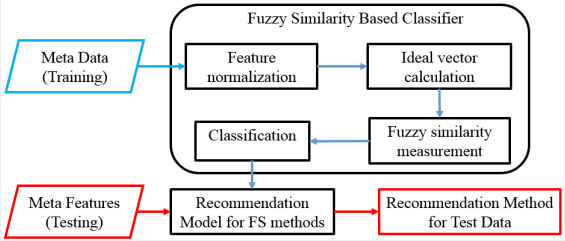


Figure 6: Framework of fuzzy similarity classifer

1. For the training set, standardize each feature using the Z-score normalization process;
2. Calculate the ideal vector $\vec{v}_l$ for the l^{th} class using geometric mean;

$$\vec{v}_l(p) = \sqrt[Z_l]{\prod_{q=1}^{Z_l} \vec{x_q}(p)},\ 1 \leq p \leq 6$$

3. Apply the same standardization process to the meta features extracted from the test dataset;
4. Implement a similarity measurement in the form of generalized Łukasiewicz algebra. Combine the similarity measures from different features using geometric mean;

$$S\langle \vec{y_r}, \vec{v}_l \rangle = \sqrt[6]{\prod_{p=1}^{6} \sqrt{1 - |\vec{y_r}(p)^2 - \vec{v}_l(p)^2|}} \quad (5)$$

5. Classify test datasets into the class of corresponding ideal vector with the highest fuzzy similarity value.

Presentation Outline

Datasets

Training Data Repository

- 1000 datasets were generated by using randomly selected parameter values.

Testing Data Repository

- 8 binary classification datasets from UCI machine learning repository.

Table 3: Parameters for data synthesis on Madelon dataset

Alias	Meaning	Value Range
P1	Number of Classes	2
P2	Number of Useful Features	[4, 5,..., 20]
P3	Number of Redundant Features	[0, 1,..., 20]
P4	Number of Repeated Features	[0, 1,..., 20]
P5	Number of Useless Features	[0, 1,..., 20]
P6	Number of Samples per Cluster	[10, 11,..., 70]
P7	Number of Cluster per Class	[2, 3,..., 7]
P8	Random Seed	[1, 2,..., 1000]
P9	Factor multiplying the hypercube dimension	[2, 3,..., 10]
P10	Fraction of y labels to be randomly exchanged	[0.01, 0.02, ..., 0.1]
P11	Flag to enable or disable random permutations	[0, 1]

Table 4: Description of the biomedical datasets for testing

Dataset	#Fea.	#Samples	Distribution Over Class
Appendicitis	7	106	85 / 21
PIMA	8	768	500 / 268
WBC	9	699	458 / 241
Statlog Heart	13	270	150 / 120
Parkinsons	22	195	48 / 147
WDBC	30	569	212 / 357
Spectfheart	44	267	55 / 212
Sonar	60	208	97 / 111

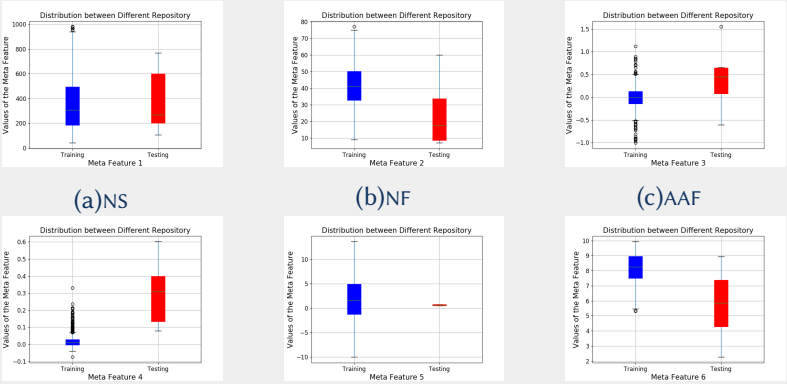
Feature Selection Methods

- Four filter FS methods which come from different categories were implemented:
  - *Statistical Based FS Method:* **Gini Index FS (GIFS)**;
  - *Similarity Based FS Method:* **ReliefF**;
  - *Information Based FS Method:* **Mutual Information FS (MIFS)**;
  - *Graph Based FS Method:* **Infinite FS (IFS)**.

- **Logistic regression** was used to evaluate the algorithms' classification performance using the generated feature rankings by different FS methods.

- The number of best performances achieved by each FS method was 546, 196, 147, 111 respectively (total of 1000 datasets).

Comparison of the Features' Distribution



(a)NS (b)NF (c)AAF

(d)ACF (e)ACVF (f)AEF

Figure 7: Comparison of the distribution between the training and testing repository

- The distributions in the training repository cover the value range of test datasets well for meta features NS, NF, ACVF;
- Meta feature AAF, ACF, AEF of the test datasets are slightly higher or lower than the corresponding value ranges.

Table 5: Weighted accuracies on 8 test datasets

Datasets	GIFS	ReliefF	MIFS	IFS	Best Method	Recommend Method
Appendicitis	**2.41**	2.40	**2.41**	2.40	G/MIFS	MIFS
PIMA	$2.63'$	2.62	**2.64**	2.56	MIFS	MIFS
WBC	$3.78'$	3.78	**3.79**	3.78	MIFS	MIFS
Statlog Heart	$4.84'$	4.61	**4.85**	4.08	MIFS	MIFS
Parkinsons	**8.52**	8.29	$8.49'$	8.42	GIFS	ReliefF
WDBC	13.48	**13.60**	$13.51'$	13.41	ReliefF	ReliefF
Spectfheart	16.03	$16.07'$	15.85	**16.11**	IFS	MIFS
Sonar	**16.07**	$15.96'$	15.31	12.83	GIFS	ReliefF

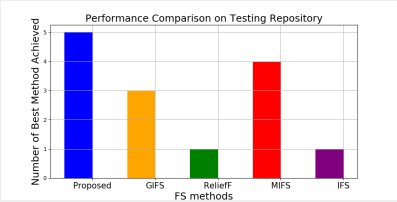


Figure 8: Performance comparison on testing repository

- Successfully recommend best method for Appendicitis, PIMA, WBC, Statlog Heart and WDBC dataset;
- Achieve the best performance comparing with the other individual FS methods;
- Overall, the successful recommendation rate was 62.5% on the testing repository.

Computational Cost

Table 6: Average run time using different methods (/s)

Datasets	Individual Methods				Meta Learning	Total Run Time
	GIFS	ReliefF	MI	IFS		
Appendicitis	0.37	0.36	0.32	0.09	0.07	0.39
PIMA	1.55	7.50	0.99	0.28	0.32	1.31
WBC	2.71	10.19	2.61	2.68	1.30	3.91
Statlog Heart	0.56	1.30	5.66	6.19	0.08	5.74
Parkinsons	2.55	1.04	1.21	0.43	0.38	1.42
WDBC	15.06	6.23	3.97	1.86	2.03	8.26
Spectfheart	3.40	3.16	4.18	2.23	0.20	4.37
Sonar	7.66	1.74	3.36	1.29	0.72	2.46

- Meta learning framework takes less than 1 second to run in most cases;
- There is comparatively little additional computational cost incurred in implementing our meta learning framework;
- The use of our meta learning method provides an efficient way to learn the potentially optimal FS method.

Presentation Outline

Discussion

- Achieve the best performance comparing with the other individual FS methods;
- Introduce very small additional computational burden and be an attractive potential to be used when a wide variety of candidate algorithms are considered;
- The results are still provisional and clearly need to be improved in the future.

Future Work

- Generate better training data repository with wider distributions, introduce more meta features and test other evaluation metrics for FS;
- Use different machine learning methods, various FS methods, datasets with diverse performance, recent meta learning models and etc.

Thank you!

Questions?

SPONSORS:

IEEE Advancing Technology for Humanity | IEEE Computational Intelligence Society | IET | The International Neural Network Society (INNS) | EPS Evolutionary Programming Society