Knowledge-Based Nonlinear to Linear Dataset Transformation for Chronic Illness Prediction

Markian Jaworsky

Co-Authors:
(1) Xiaohui Tao, Ji Zhang
(2) Jianming Yong
(3) Lei Pan, Shiva Pokhrel

1: School of Mathematics, Physics, and Computing, University of Southern Queensland, Toowoomba
2: School of Business, University of Southern Queensland, Springfield
3: School of Information Technology, Deakin University, Australia

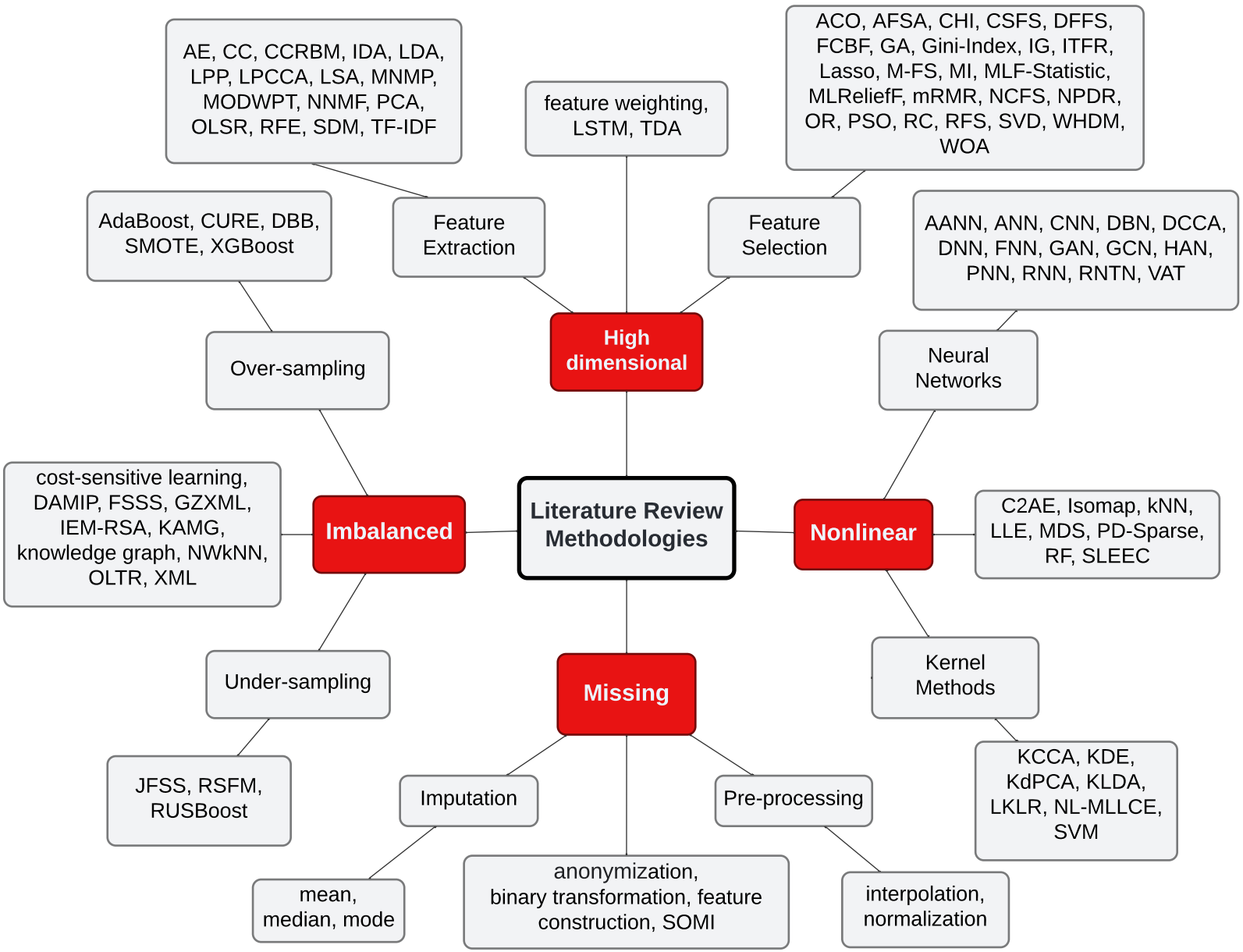
Introduction

- Predictive models trained on nonlinear data are prone to overfitting and generally have poorer performance in classifying unseen data. [1]
- Spurious correlations are likely to occur with nonlinear patterns and not be fully explainable. [2]
- Linear correlations are explainable with traditional statistical methods.

Literature Review - Data Challenges

- High Dimensional Datasets - 38% review citations
- Imbalanced Class Datasets - 24% review citations
- Missing and Erroneous Data - 22% review citations
- Multiple and Nonlinear Datasets - 16% review citations

Literature Review - Method Complexity



Research Concept - Knowledge Graphs

- We leverage a knowledge graph to transform health survey responses, which consist of a majority of nonlinear nominal variables, into meaningful and similarly scaled linear values.
- The linear values give a more meaningful pattern to both human interpretation and improve the Multinomial Naive Bayes classifier.

Knowledge Graph - Automation

- We re-use the knowledge graph generated via automation as we presented at last year's 2022 HIS conference [3].
- Feature relations are determined by significant rankings of term frequencies in the WHO ICD Code chapters of human organ systems.

https://github.com/mjaworsky/KnowledgeGraph

Research Design - Datasets

- www.cdc.gov / brfss / annual_data
- www.who.int / standards / classifications / classification-of-diseases

Research Design - Overview

- Our knowledge-based linear variable transformation is derived using a corpus of key terms associated with cancer and diabetes and cross-referencing each term against the 26 WHO ICD chapters.

Health Survey Keyword to Linear Variable Output

ICD Chapter	male	female	diabetes	cancer	obese	overweight	smoked	cigarettes	sugar
1	0	3	2	3	0	0	0	0	0
2	17	21	0	39	0	0	0	0	0
3	0	0	0	2	0	0	0	0	0
4	1	1	0	2	0	0	0	0	0
5	4	4	89	6	0	25	0	0	3
6	0	0	1	3	0	0	3	2	0
7	0	0	2	0	0	0	0	0	0
8	1	0	9	10	0	1	0	0	0
9	0	0	10	3	1	0	0	0	1
10	0	0	1	0	0	0	0	0	0
11	1	0	2	1	0	0	0	0	0
12	0	0	2	0	0	0	0	0	1
13	0	2	0	2	0	0	0	0	0
14	11	6	10	5	2	0	0	0	0
15	0	1	2	2	0	0	0	0	0
16	26	86	7	7	0	0	0	0	0
17	30	13	0	0	0	0	0	0	0
18	0	5	10	0	0	0	0	0	0
19	0	0	20	0	0	1	0	0	0
20	17	13	1	3	0	0	0	0	0
21	28	29	4	36	1	0	0	0	1
22	1	4	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0
24	1	3	2	1	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0
26	2	7	0	0	0	0	0	0	0

Research Evaluation - Classifiers

- AB = AdaBoost
- CNN = Convolutional Neural Network
- KNN = K-Nearest Neighbours
- LDA = Linear Discriminant Analysis
- LR = Logistic Regression
- MNB = Multinomial Naive Bayes
- RB = RUSBoost
- RF = Random Forest
- SVM = Support Vector Machine

Baseline v Feature Selection v Linear Transformation Results

2*Dataset	2*Classifier	Baseline			Feature Selection			Linear Variable		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
9*2017-2021	AB	0.51	0.50	0.48	0.67	0.51	0.52	0.43	0.50	0.46
	CNN	0.26	0.52	0.29	0.41	0.50	0.40	0.37	0.50	0.37
	KNN	0.47	0.50	0.47	0.50	0.50	0.48	0.43	0.50	0.46
	LDA	0.49	0.50	0.48	**0.81**	0.61	0.58	0.43	0.48	0.45
	LR	0.48	0.50	0.47	0.66	0.50	0.47	0.43	0.50	0.46
	MNB	0.48	0.51	0.44	0.47	0.51	0.47	**0.81**	**0.77**	**0.74**
	RB	0.50	0.50	0.49	0.73	0.68	0.67	0.46	0.49	0.46
	RF	0.46	0.50	0.48	0.56	0.50	0.47	0.46	0.50	0.46
	SVM	0.46	0.50	0.48	0.46	0.50	0.48	0.46	0.50	0.46

Conclusion 1

- Knowledge graphs can be used to transform nonlinear variables into a set of new linear variables.

Conclusion 2

- A dataset of knowledge-based linear-only variables consistently outperformed a baseline dataset with many thresholds of feature selection including a neural network solution.

USQ Human Research Ethics Approval

- USQ HREC ID: H21REA222
- Approval date: 15/10/2021
- Reviewed date: 13/10/2022
- Expiry date: 15/10/2024
- USQ HREC status: Approved

References

[1] Louis Létinier et al. **“Artificial intelligence for unstructured healthcare data: application to coding of patient reporting of adverse drug reactions”.** In: *Clinical Pharmacology & Therapeutics* (2021). DOI: *10.1002/cpt.2266*. URL: *https://doi.org/10.1002/cpt.2266*.

[2] Peter Washington et al. **“Challenges and opportunities for machine learning classification of behavior and mental state from images”.** In: *arXiv preprint arXiv:2201.11197* (2022).

[3] Markian Jaworsky et al. **“Automated Knowledge Graph Construction for Healthcare Domain”.** In: *International Conference on Health Information Science*. Springer. 2022, pp. 258–265.

Thanks

Questions?

Markian Jaworsky
Markian.Jaworsky@usq.edu.au