An Empirical Study on the Robustness of Knowledge Injection Techniques Against Data Degradation

Andrea Rafanelli[a,b] Matteo Magnini[c] Andrea Agiollo[c]
Giovanni Ciatto[c] Andrea Omicini[c]

[a]Dipartimento di Informatica – Università di Pisa

[b]Dipartimento di Informatica – Scienza e Ingegneria e Matematica – Università dell'Aquila
andrea.rafanelli@phd.unipi.it

[c]Dipartimento di Informatica – Scienza e Ingegneria
Alma Mater Studiorum – Università di Bologna
{matteo.magnini, andrea.agiollo, giovanni.ciatto, andrea.omicini}@unibo.it

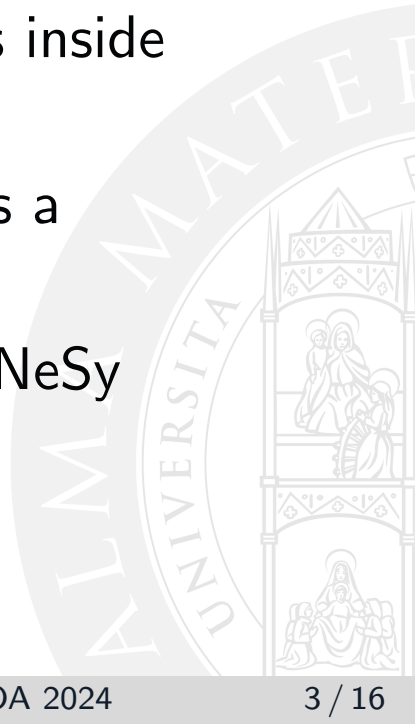
25th Workshop "From Objects to Agents"
8th-10th July, 2024, Forte di Bard (Italy)

Outline

Next in Line...

SKI in Intelligent Agents

- sub-symbolic approaches – e.g., neural networks (NNs) – enabled computational agents with smart behaviours such as speech recognition [8], object detection [10] and more
- autonomous agents' knowledge represented *symbolically* both at the conceptual and technological level
- requirement for integration of *sub*-symbolic components inside autonomous intelligent agents
- *neuro-symbolic* integration (NeSy) in agents emerged as a solution [2, 3]
- *symbolic knowledge injection* (SKI) [4] most promising NeSy technique for agents



Background on SKI

SKI definition

any *algorithmic* procedure affecting how sub-symbolic predictors draw their inferences in such a way that predictions are either *computed* as a function of, or made *consistent* with, some *given* symbolic knowledge

- symbolic knowledge represent agent's beliefs, well-known concepts, societal norm or any desirable rule for the sub-symbolic system to be considered
- symbolic knowledge as logic rules, knowledge graphs, expert knowledge, etc.
- SKI attained by modifying the NN structure, altering the training process, or extending the training data to take into account the symbolic knowledge

Advantages and issues of SKI

SKI advantages

- less data required to train (gather information from injected knowledge)
- less time required to train (fewer concepts to learn from data)
- more trustworthy (outcome following injected knowledge)
- more robust to data corruption (repair information using injected knowledge)

Current lack in SKI

- measuring robustness of injection mechanisms is crucial for developing neuro-symbolic agents
- lack of trustworthiness measurements in SKI [1]

Paper context

Idea

focus on SKI, present comprehensive modelling of new *robustness* metric

Robustness definition

resilience of SKI training over imperfect, bugged or missing data

Next in Line...

Robustness definition

Key idea

injection is *robust* if predictive performance of educated predictor is poorly affected by *perturbations* of training data, as long as small perturbation *magnitude*

Data perturbation

altering training dataset D by **adding**, **removing**, or **editing** its entries

- denote the perturbed dataset as $D' = D \circ \Delta D$
- denote by $\|\Delta D\| \in \mathbb{R}_{\geq 0}$ the *magnitude of the perturbation*

Robustness score definition

Robustness score

$$\rho_{N,D}(\mathbf{D}) = \frac{1}{n} \sum_{\Delta D \in \mathbf{D}} \|\Delta D\| \cdot \frac{\pi(N_{\Delta D}, D \circ \Delta D)}{\pi(N, D)} \tag{1}$$

- $\mathbf{D} = \{\Delta D_1, \ldots, \Delta D_n\}$ set of data perturbations applied to D
- N the predictor trained on D
- $N_{\Delta D}$ the same predictor trained on the perturbed dataset $D \circ \Delta D$
- π is a performance metric of choice, such as accuracy

SKI Robustness gain definition

Injection robustness measured by applying Equation (1) to educated predictor $\hat{N} = \mathcal{I}(N, K, D)$, attained injecting the knowledge K, on some uneducated predictor N, then trained upon D

Robustness gain

$$R_{N,D}(\mathcal{I}) = \frac{\rho_{\hat{N},D}(\mathbf{D})}{\rho_{N,D}(\mathbf{D})} \tag{2}$$

- $R_{N,D}(\mathcal{I}) > 1$ indicates the injection mechanism $\mathcal{I}$ produces a more robust predictor
- injection mechanisms suffering data perturbations result in $R_{N,D}(\mathcal{I}) < 1$
- $R_{N,D}(\mathcal{I})$ is easy-to-understand measure for analysing robustness quality of a SKI mechanism

Measuring data perturbations

- $\rho_{N,D}(\mathbf{D})$ requires measuring the magnitude $\|\Delta D\|$
- measure $\|\Delta D\|$ by measuring the *difference* among two datasets A, B
- leverage on the Kullback-Leibler (KL) divergence [5]

$$\psi(A,B) = \frac{1}{2}\Big[\mathrm{tr}(\boldsymbol{\sigma}_B^{-1}\boldsymbol{\sigma}_A) - \mathrm{dim}(A) + \ln\left(\frac{\det\boldsymbol{\sigma}_B}{\det\boldsymbol{\sigma}_A}\right) + (\boldsymbol{\mu}_B - \boldsymbol{\mu}_A)^\top \boldsymbol{\sigma}_B^{-1}(\boldsymbol{\mu}_B - \boldsymbol{\mu}_A)\Big]$$

- KL divergence computed per-class $A = A_1 \cup \ldots \cup A_K$ and $B = B_1 \cup \ldots \cup B_K$, then

$$\Psi(A,B) = \frac{1}{|A|}\sum_{k=1}^{K} \psi(A_k, B_k) \cdot |A_k|$$

Data perturbation measure

$$\|\Delta D\| = \Psi(D, D') \tag{3}$$

Next in Line...

Experiments setup I

Datasets

- **Breast Cancer**: BCW contain 699 instances of breast cancer clinical exams with benign and malignant classes
- **Splice Junction**: PSJGS dataset contain 3,190 gene sequence instances and has three classes
- **Census Income**: CI dataset contain 48,842 instances representing individuals with binary classes representing person income

SKI methods

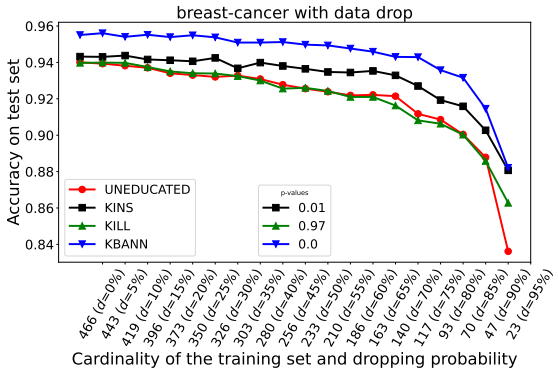
- **KINS**, Knowledge Injection via Network Structuring [6]
- **KILL**, Knowledge Injection via Lambda Layer [7]
- **KBANN**, Knowledge-Based Artificial Neural Network [9]

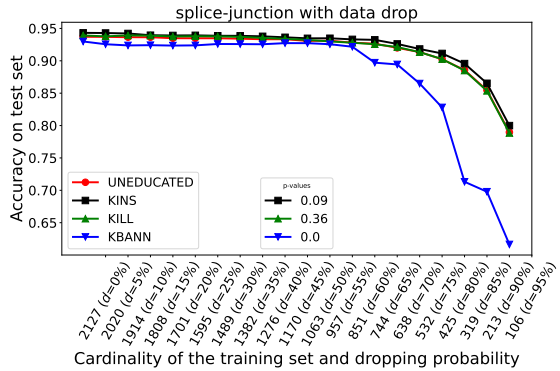
Experiments setup II

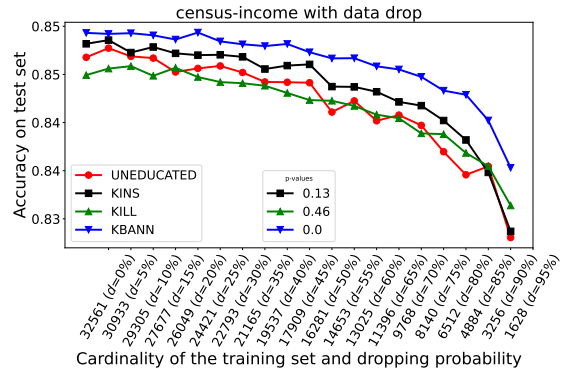
Data perturbation strategies

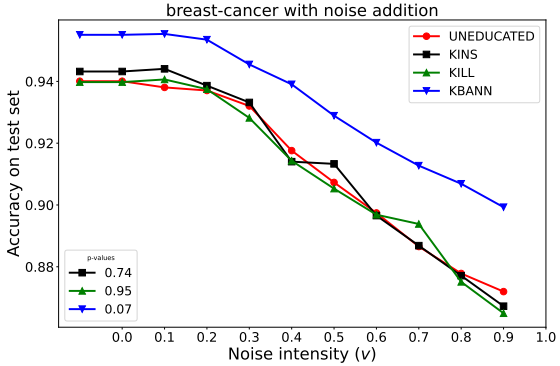
- **Sample Drop**: mimics the effect of an intelligent agent lacking training data. Aims at selectively mutilating a dataset.
- **Noise Addition**: mimics situation where data acquisition process of the autonomous agent is affected by error. Aims at degrading a dataset in a *controlled* way.
- **Label flipping**: mimics situation where data labelling process is affected by error. Selectively flipping labels of some entries in a dataset.

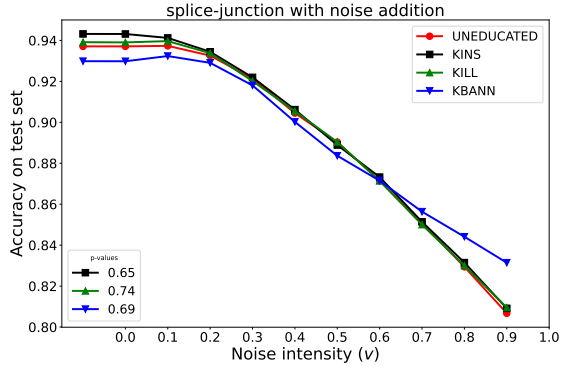
Experiments results I

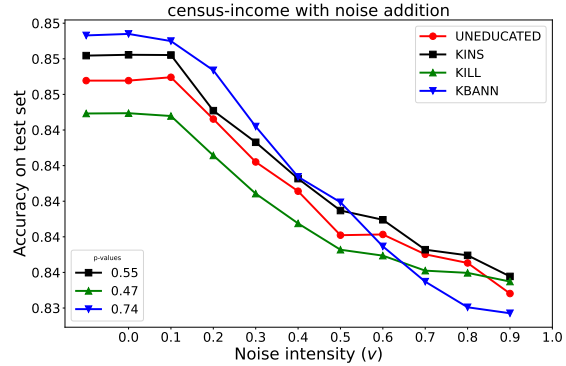


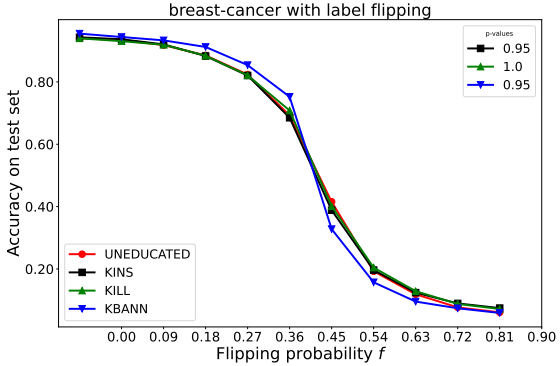


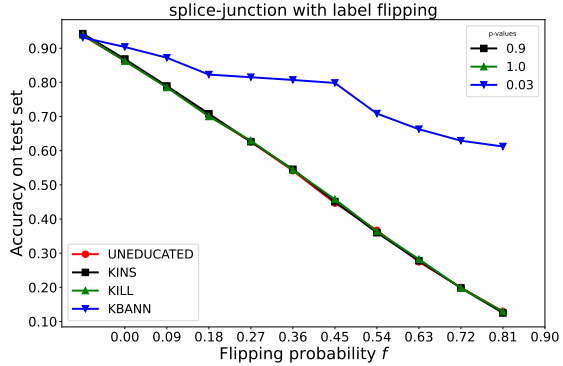


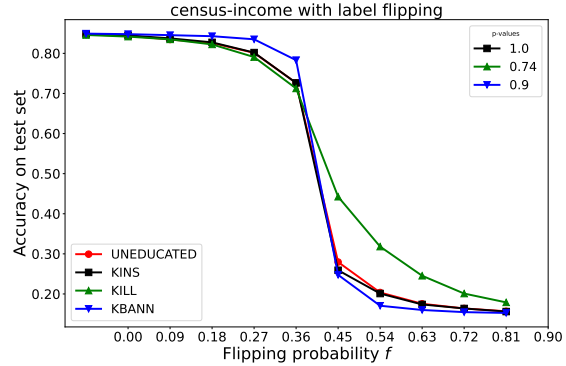












Experiments results II

Dataset	$R_{N,D}(\mathcal{I})$ drop			$R_{N,D}(\mathcal{I})$ noise			$R_{N,D}(\mathcal{I})$ flip		
	KINS	KILL	KBANN	KINS	KILL	KBANN	KINS	KILL	KBANN
BCW	**1.0493**	**1.0318**	**1.0382**	0.9960	0.9985	**1.0109**	0.9994	**1.0184**	0.9520
PSJGS	**1.0045**	0.9968	0.8425	0.9950	0.9984	**1.0145**	0.9962	**1.0026**	**1.6749**
CI	0.9998	**1.0039**	**1.0043**	0.9992	**1.0012**	0.9965	0.9897	**1.1703**	0.9815

Findings

- SKI heightened robustness when data is limited (alternate guidance of integrated knowledge)
- loss-manipulating SKI better tolerate label corruptions (deemphasise flawed labels during backpropagation)
- SKI structuring methods more robust than constrained-layer types (injected knowledge kept intact)

Next in Line...

Summing up

Contributions

- propose novel metric for robustness of SKI
- define three data perturbation strategies and perturbation metric
- robustness gain measure to assess whether SKI is better than its uneducated counterpart
- experimental evaluation of our robustness score metric over different datasets, injectors and perturbations
- showcase SKI robustness for label corruption

An Empirical Study on the Robustness of Knowledge Injection Techniques Against Data Degradation

Andrea Rafanelli[a,b] Matteo Magnini[c] Andrea Agiollo[c]
Giovanni Ciatto[c] Andrea Omicini[c]

[a]Dipartimento di Informatica – Università di Pisa

[b]Dipartimento di Informatica – Scienza e Ingegneria e Matematica – Università dell'Aquila
andrea.rafanelli@phd.unipi.it

[c]Dipartimento di Informatica – Scienza e Ingegneria
Alma Mater Studiorum – Università di Bologna
{matteo.magnini, andrea.agiollo, giovanni.ciatto, andrea.omicini}@unibo.it

25th Workshop "From Objects to Agents"
8th-10th July, 2024, Forte di Bard (Italy)

References I

[1] Andrea Agiollo and Andrea Omicini.
Measuring trustworthiness in neuro-symbolic integration.
In *Proceedings of the 18th Conference on Computer Science and Intelligence Systems*, volume 35 of *Annals of Computer Sciences and Information Systems*, pages 1–10, September 2023.

[2] Andrea Agiollo, Andrea Rafanelli, Matteo Magnini, Giovanni Ciatto, and Andrea Omicini.
Symbolic knowledge injection meets intelligent agents: QoS metrics and experiments.
Autonomous Agents and Multi-Agent Systems, 37(2):27:1–27:30, June 2023.

[3] Andrea Agiollo, Andrea Rafanelli, and Andrea Omicini.
Towards quality-of-service metrics for symbolic knowledge injection.
In Angelo Ferrando and Viviana Mascardi, editors, *WOA 2022 – 23rd Workshop "From Objects to Agents"*, volume 3261 of *CEUR Workshop Proceedings*, pages 30–47, November 2022.

[4] Giovanni Ciatto, Federico Sabbatini, Andrea Agiollo, Matteo Magnini, and Andrea Omicini.
Symbolic knowledge extraction and injection with sub-symbolic predictors: A systematic literature review.
ACM Computing Surveys, 56(6):161:1–161:35, June 2024.

[5] James M. Joyce.
Kullback-Leibler divergence.
In Miodrag Lovric, editor, *International Encyclopedia of Statistical Science*, pages 720–722. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[6] Matteo Magnini, Giovanni Ciatto, and Andrea Omicini.
KINS: Knowledge injection via network structuring.
In Roberta Calegari, Giovanni Ciatto, and Andrea Omicini, editors, *CILC 2022 – Italian Conference on Computational Logic*, volume 3204 of *CEUR Workshop Proceedings*, pages 254–267, 2022.

References II

[7] Matteo Magnini, Giovanni Ciatto, and Andrea Omicini.
A view to a KILL: Knowledge injection via lambda layer.
In Angelo Ferrando and Viviana Mascardi, editors, *WOA 2022 – 23rd Workshop "From Objects to Agents"*, volume 3261 of *CEUR Workshop Proceedings*, pages 61–76. Sun SITE Central Europe, RWTH Aachen University, November 2022.

[8] Ali Bou Nassif, Ismail Shahin, Imtinan Basem Attili, Mohammad Azzeh, and Khaled Shaalan.
Speech Recognition Using Deep Neural Networks: A Systematic Review.
IEEE Access, 7:19143–19165, 2019.

[9] Geoffrey G. Towell, Jude W. Shavlik, and Michiel O. Noordewier.
Refinement of approximate domain theories by knowledge-based neural networks.
Proceedings of the 8th National Conference on Artificial Intelligence, pages 861–866, 1990.

[10] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu.
Object Detection With Deep Learning: A Review.
IEEE Transactions on Neural Networks and Learning Systems, 30(11):3212–3232, 2019.