WITAN: Unsupervised Labelling Function Generation for Assisted Data Programming

Benjamin Denham, Edmund M-K Lai, Roopak Sinha, and M. Asif Naeem

Funded by a Callaghan Innovation R&D Fellowship Grant FPAP1902 for Fisher & Paykel Appliances and a Doctoral Fees Scholarship from Auckland University of Technology

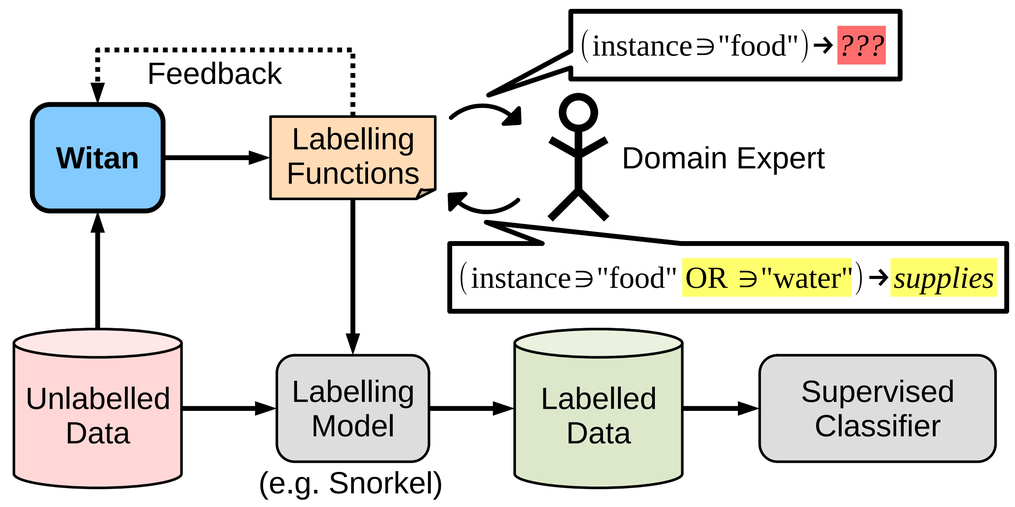
Motivation

- Labelling training data is often **prohibitively expensive**
- **Data Programming** provides cheaper **weak supervision** in the form of **labelling functions (LFs)**
- But: users must still **manually craft** LFs that assign class labels
- Even **assisted data programming** requires non-trivial effort:

Does not require:	Labelled instances	Designed LFs	Prior set of classes	Continuous feedback
LF design aids	✔	✘	✘	✘
Instance-based LF design aids	✘	✘	✘	✔
Instance-based LF generation	✘	✔	✘	✔
Feedback-based LF generation	✔	✔	✘	✘
LFs from clusters	✔	User interaction not supported		
Clustering by intent	Min. 1 class	✔	✔	✔
Our contribution: WITAN	✔	✔	✔	✔

Contribution

- **WITAN** proposes LF conditions **without any initial supervision**
- Users **assign classes** to LFs
- Users may **edit LF conditions**
- WITAN can learn from **optional user feedback**



WITAN Algorithm

Here we demonstrate using WITAN to **categorise aid requests after a disaster**, where identifying the class set and classifier training are both **time critical**

1. Start with a set of **candidate LF conditions** constructed from binary features (e.g. bag-of-words)

	∋ help	∋ need	∋ food	∋ like	∋ know	∋ water	∋ information	...
help	0	0	0	25	28	0	125	...
need	0	0	0	191	11	0	0	...
food	0	0	0	85	45	0	116	...
like	27	184	84	0	0	78	0	...
know	29	12	43	0	0	38	0	...
water	0	0	0	72	36	0	52	...
information	104	0	95	0	0	48	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱
$U =$	**253**	**373**	**611**	**594**	**248**	**408**	**693**	**...**

2. Construct a **matrix of the information gain** each LF condition provides on each feature for instances matched by the condition

3. Each LF condition's **Utility (U)** is the **sum of gain over all features**

4. The LF condition with **highest utility** is selected next to propose to the user. E.g. They may assign a class *advice* to complete the LF: ∋ information → *advice*

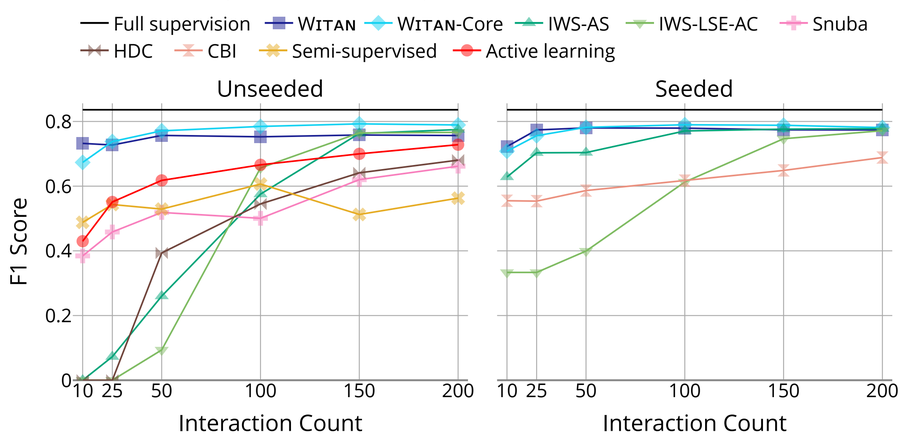
5. The selected condition is removed, **all cells are updated** to account for its information gain, and the process repeats

- WITAN resembles supervised rule learners that **maximise information gain on the class variable**
  - As class values are unknown, WITAN assumes that **some features correlate with the class**
  - As class-correlated features are unknown, WITAN **maximises information gain across all features**
- 3 extensions to the core algorithm:
  - Candidate **AND-conditions** can be generated from conjunctions of selected LFs and other features
  - Generating higher-utility **OR-conditions** from disjunctions of logically similar conditions
  - **User feedback** on conditions can weight the matrix for subsequent selections of conditions

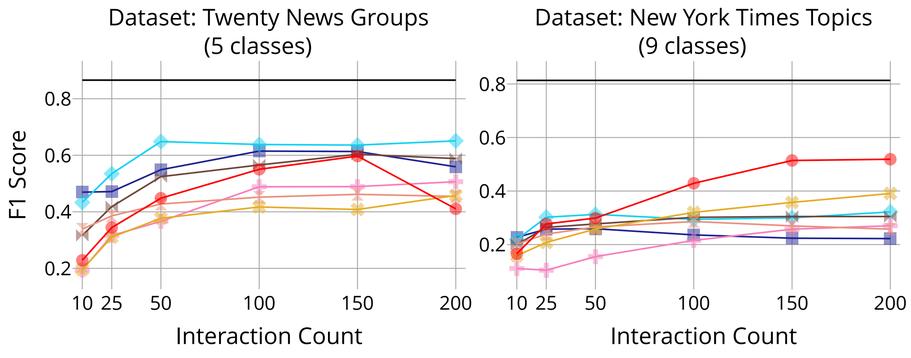
Results

- Setup to compare labelling methods:
  - LF-based methods use the **Snorkel** labelling model
  - **User interactions simulated** from ground truth labels
- For binary classification tasks:
  - WITAN is the **top performer** on **most datasets**
  - WITAN reaches peak performance with **fewer interactions**, even without feedback
- For multi-class classification tasks:
  - WITAN is the **top performer** on **some datasets**
  - **instance-level labelling** methods perform better on datasets with **many classes**
- WITAN is **up to 70× faster** than all other LF generators, except on high feature counts

Binary Classification (Movie Review Sentiment)

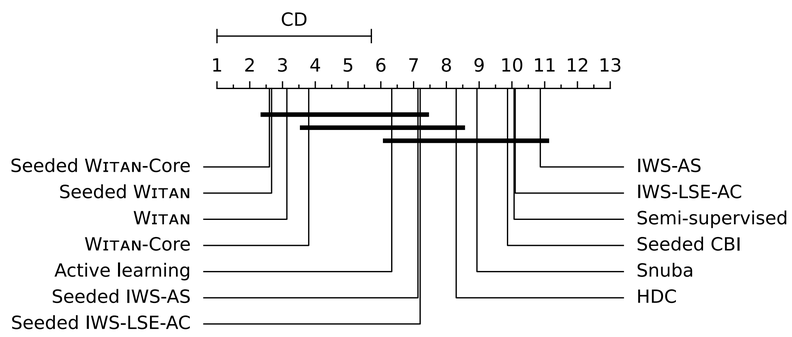


Multi-Class Classification

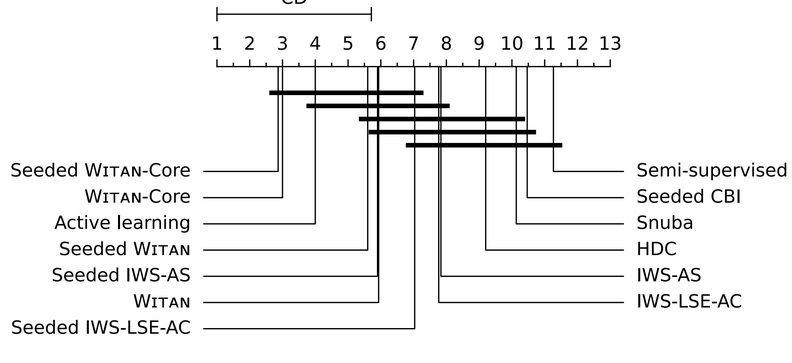


WITAN is **overall top performer** for binary classification:

After 25 interactions



After 100 interactions



Conclusions

- WITAN supports **more interaction modes** than prior assisted data programming:
  - Unsupervised — users do not need to specify a set of classes up front
  - Users do not need to provide continuous feedback to drive LF generation
- WITAN achieves **competitive performance** in binary and multi-class tasks **without initial supervision**

Future Work

- Applying WITAN to features derived from other data types (e.g. **numeric or image** data)
- Leveraging **instance-level active learning** to complement coarser LFs

AUT
School of Engineering, Computer and Mathematical Sciences
Auckland University of Technology
ben.denham@aut.ac.nz
github.com/ben-denham/witan