Final Presentation

EML4U

October 28, 2022

Specific Project Details & Results

- Drift detection & Explanations
  Universität Bielefeld Institut für Kognition und Robotik (CoR-Lab)
  Robert Feldhans
- Drift Explanations via Polygons and Hyperboxes
  Universität Paderborn - Arbeitsgruppe Data Science (DICE)
  Adrian Wilke
- Uncertainty Quantification
  Universität Paderborn - AG Intelligente Systeme und Maschinelles Lernen (ISML)
  Mohammad Hossein Shaker
- Semalytix' Business Use Case
  Semalytix GmbH
  Fabian Hommel

Drift detection & Explanations

- Joint Paper *Drift detection in text data with document embeddings* [1] at IDEAL
- Drift detection is crucial for drift explanation
- Tested four drift detectors with two datasets in several scenarios
- Least-Squares Density Difference and Kernel-Two-Sample best Drift detectors, LSDD better on real-world Twitter dataset
- Lower embedding dimensions tend to produce better drift detection results

Drift explanation via Difference Accentuation [2]

- Drift detected means new distribution differs from old one
- Accentuate these differences by creating many dimension reductions and choose the one where both distributions differ the most
- Then use this dimension reduction to cluster the data; differentiate each cluster via labels (tf-idf)

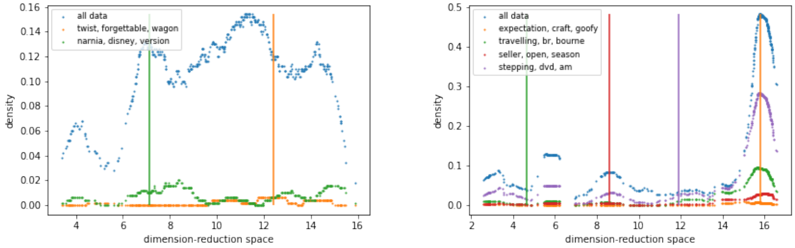


Figure: Data from the Amazon movie review dataset with one (left) and five star reviews (right).

Model update explanation via Contrastive Explanations [3]

- Contrastive explanations take a data point and calculate the most similar one of a different classification
- Do this with a lot of points before and after model update and calculate impact for each parameter

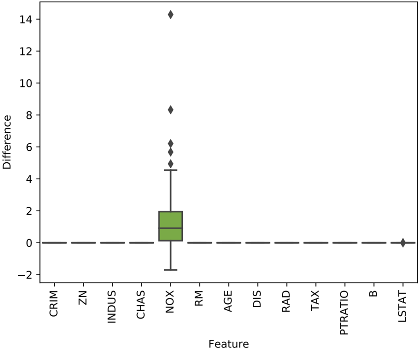


Figure: Changes in counterfactual explanations for the house prices data set.

Drift Explanations via Polygons and Hyperboxes

- Article (in progress):
  Explaining Drift in Text Data with Document Embeddings [4]
- Bridging the gap:
  - Unsupervised approaches, no labels
  - Benchmark dataset to evaluate approaches
- Drift Explanation with Polygons and with Hyperboxes
- Use resulting drift explanations to resolve model conflicts

Drift Explanation via Polygons

Text

"Might have been profound for it's time but... When I pay almost 30. for a DVD I expect alot ..."

↓ **Embed**
BERT / Doc2Vec (BoW)

High-dimensional embeddings (e.g. 50 or 768 dimensions)

[-0.02514367, -0.29414916, ... 0.40210924, -0.09406912]

↓ **Reduce**
UMAP / TSNE / PCA

2-dimensional embeddings

[0.04848728, -0.2670002]

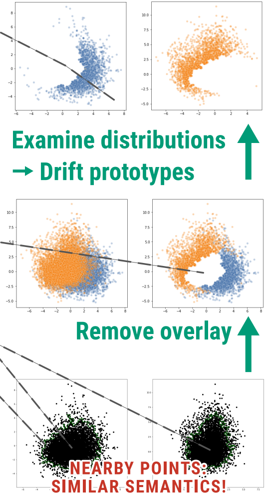
Detect polygons →

Clusters?
→ multiple
explanations

Overlay:
Similar semantics
in distributions
(uncertainty)

Drift expected
in distributions

Shrink → Adjust



Examine distributions
→ Drift prototypes ↑

Remove overlay ↑

Drift Explanation via Hyperboxes

Setup: Two sets of embeddings, A and B.

Step 1: Collect values of single dimensions

A_1: `[2,6,3,8,5,...]`
A_2: `[7,4,1,0,9,...]`
$Dim_1(A)$: `[2,7,...]`
$Dim_2(A)$: `[6,4,...]`

Step 2: Create 1-dimensional bounding box for each dimension. Remove outliers (percentiles).

$Dim_1(A)$: `[2,7,7,6,7,9,9,42,9]`
$Dim_2(A)$: `[6,4,1,6,4,6,4,55,6]`
$\rightarrow Box_{Min}(A_1) = 6, Box_{Max}(A_1) = 9$
$\rightarrow Box_{Min}(A_2) = 4, Box_{Max}(A_2) = 6$

Step 3: Get prototypes by checking if values of embeddings B are inside bounding boxes of A.

B_1: `[8,3,...]`
$B_1 \in Box(A_1)$? $6 \leq 8 \leq 9 \rightarrow$ Yes
$B_1 \in Box(A_2)$? $4 \not\leq 3 \leq 6 \rightarrow$ No
Score = 1 + 0 + ... < total dimensions
$\rightarrow B_1 \in Prototypes$

Result: Frequent Words in Clusters

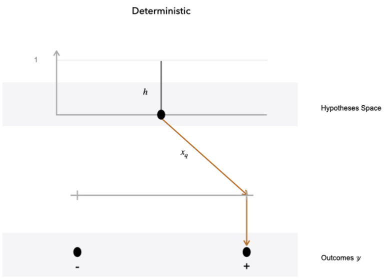
(Words in multiple clusters removed, e.g. movie, film, dvd)

Uncertainty Quantification

Representing uncertainty in ML

Ensemble-based Uncertainty Quantification: Bayesian versus Credal Inference [5]

- Machine learning is inseparably connected with uncertainty.
- Uncertainty of a learner can be represented in different levels:
  - Level 0: Deterministic
  - Level 1: Probabilistic
  - Level 2: Bayesian

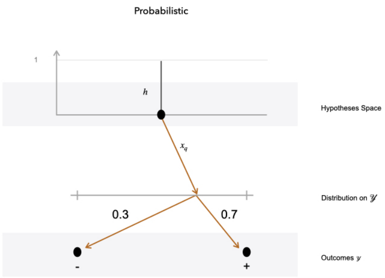


Uncertainty Quantification

Representing uncertainty in ML

Ensemble-based Uncertainty Quantification: Bayesian versus Credal Inference [5]

- Machine learning is inseparably connected with uncertainty.
- Uncertainty of a learner can be represented in different levels:
  - Level 0: Deterministic
  - Level 1: Probabilistic
  - Level 2: Bayesian

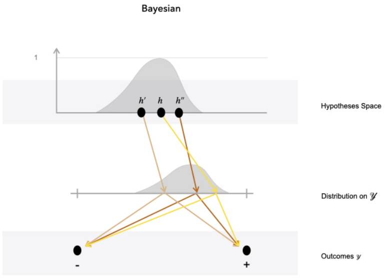


Uncertainty Quantification

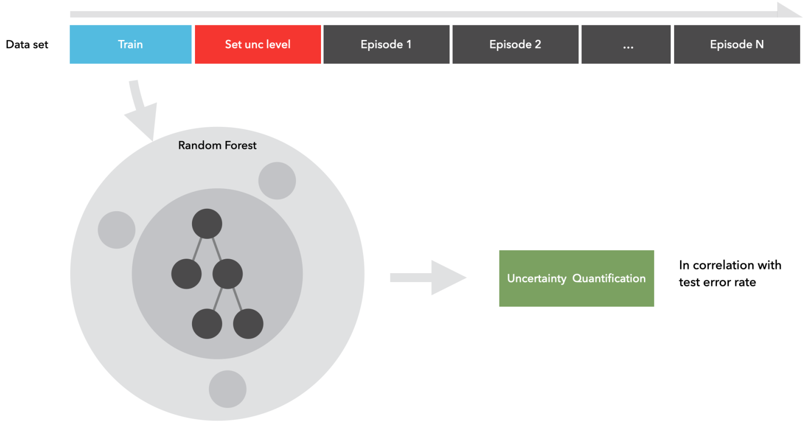
Representing uncertainty in ML

Ensemble-based Uncertainty Quantification: Bayesian versus Credal Inference [5]

- Machine learning is inseparably connected with uncertainty.
- Uncertainty of a learner can be represented in different levels:
  - Level 0: Deterministic
  - Level 1: Probabilistic
  - Level 2: Bayesian



Uncertainty drift detection



Model update strategies

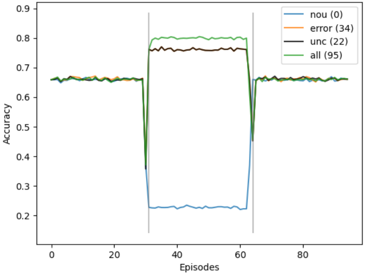
- No model update: train the model only with the first episode and never update.
- Uncertainty: Update the model only if the uncertainty value of the new episode is higher than base line uncertainty value (Only requires labels when making updates).
- Error rate: Update the model if the error rate of the new episode is higher than base line error rate (requires labels for every episode)
- Update on every episode: requires labels for every episode and the most resource intensive

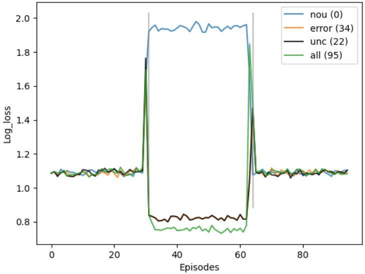
Experiment and results

Synthetic Data Generation detail

- normal_samples = 100000
- Drift samples (different distribution) = 50000
- n_features= 50
- n_informative_features= 30
- n_classes = 5

Results averages over 10 runs with different random seeds



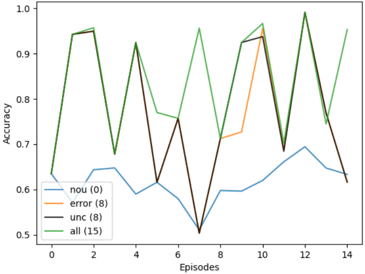


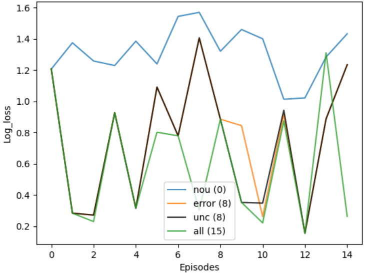
Experiment and results

Amazon movie reviews dataset

- normal_samples = 10000
- n_features= 50
- n_classes = 5

Results averages over 10 runs with different random seeds

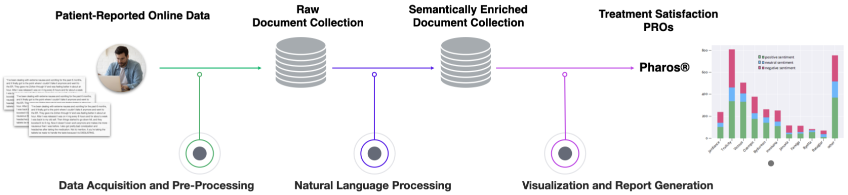




Semalytix' Business Use Case

- (1) Business Use Case and Problem Definition
- (2) Data and ML Model
- (3) Global Drift Detection
- (4) Localized Drift Detection
- (5) Interactive Case Study
- (6) Results and Discussion
- (7) Conclusion

Semalytix’ Business Use Case



Business Use Case and Problem Definition

Context:

- Semalytix is trying to understand needs and burdens of patients in online patient experience text data
- This data is very heterogeneous: Style, emotional content and level of medical expertise
- We turn this unstructured data into structured data via a large suite of NLP models
- We regularly receive data updates or completely new sources of data

Problem:

- Given a model M that was trained on a corpus D1 and a new data source D2, does M generalize to D2 without substantial loss in performance?
- Without annotated ground truth!

Experimental Procedure

For experiments, we used data from 5 real use cases

- Reference data (D1) is always from the same big training corpus
- Target data (D2) data is from five distinct sources that models have not seen
- We sample 10k documents per corpus, pre-filter by relevance for life sciences and split the documents into sentences.

Experimental scenarios:

- Global drift detection in target data
- Local drift detection in target data

Experiment Data and ML Model

- The tested model was a transformer-based medical sentiment model

	Number of Sentences in D1	Number of Sentences in D2
Corpus Pair 1	36524	24822
Corpus Pair 2	36886	11981
Corpus Pair 3	36306	11700
Corpus Pair 4	37699	11913
Corpus Pair 5	37893	11174

Table: Number of sentences per corpus pairing.

Global Drift Detection

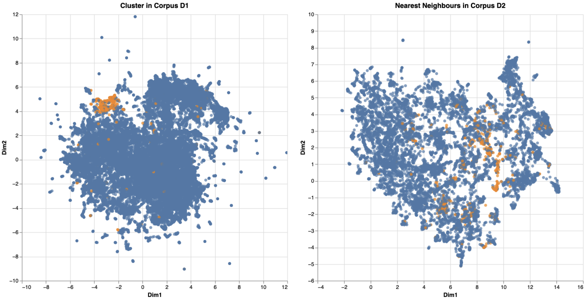
Research Question: Which methods can be used to detect drift in patient experience data?

- Two established methods [1]: the Kolmorogov-Smirnov test (KS) and the Least-Squares Density Difference Estimation method (LSDD)
- Three other, distribution-distance-based methods:
  1. Jensen-Shannon-Distance (JSD) between word count distributions
  2. JSD between predicted label distributions
  3. JSD between predicted label probability distributions

Localized Drift Detection

Research Question: Is it possible to localize regions of strong drift in the target data?

- We embed all sentences for a corpus pair D1 and D2 with a transformer model
- Then, we cluster the embedded sentences in D1
- For each cluster in D1, we compute the centroid and obtain its k nearest neighbors in D2. k is chosen as the size of the cluster in D1
- This results in a list of cluster pairings that we can examine with drift detection methods

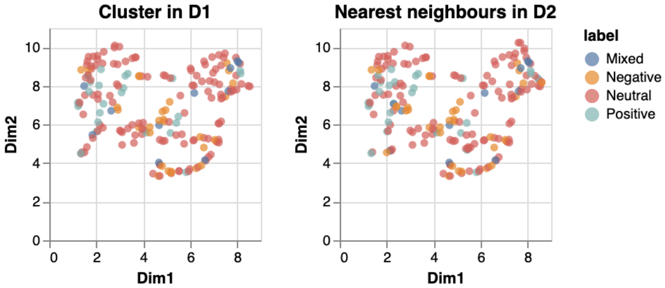


Interactive User Interface Case Study

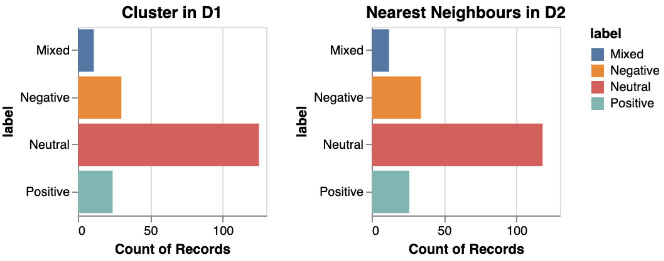
Research Question: Can the distance measurements from the localized approach be used in a user interface to decide whether a model update is necessary?

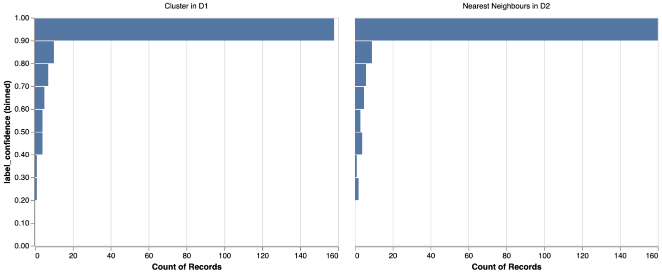
- We sort the local region pairs by descending distance
- For the region pairs with highest distance, study subjects were asked to rate the perceived difference between regions based on four comparative visualizations:
  1. Word Clouds
  2. Scatter Plots
  3. Predicted Label Bar Charts
  4. Predicted Probability Histograms

Interactive User Interface Case Study



Interactive User Interface Case Study





Results of Global and Localized Drift Detection

- Global Drift Detection
  - KS and LSDD detected **drift in all five corpus pairs**
  - **KS and LSDD correlate strongly with the Word Count Distribution Distance and also with the Label Distribution Distance**
    - might be due to the sentiment task, which is sensitive to certain words
  - The Prediction Probability Distribution Distance does not correlate with any of the other methods
    - Does active learning lead into a different direction than our approach?
- Local Drift Detection
  - **Variance in distances is much higher**
  - KS and LSDD detect **drift in some regions, but not all**

Results of the Interactive Interface Study

- Interactive Interface Case Study
  - Users report that the **interface is appealing, but it is hard to find the perceived effects in the raw data**
  - The ratings from the label bar charts and the scatter plots correlate with the respective distance metrics
    - These visualizations seem to provide meaningful information
  - The ratings for the word clouds and probability histograms did not correlate with the respective distance metrics
    - Discard or improve these visualizations

Conclusion

- Finding ground truth for drift detection in real-world patient experience data is hard!
- Future research needs to be invested into ground-truth manifestation of drift in real-world data sets
- The localized approach is promising, especially in a human-in-the-loop interface
- Semalytix has started initiatives to integrate local drift detection into productive workflows
- A lot of open opportunities for fine-tuning the approach
  - The choice of detection methods and their inputs
  - How to choose and pair regions
  - Which visualizations to use

EML4U Publications I

[1] R. Feldhans, A. Wilke, S. Heindorf, M. H. Shaker, B. Hammer, A.-C. Ngonga Ngomo, and E. Hüllermeier, "Drift Detection in Text Data with Document Embeddings," in *Intelligent Data Engineering and Automated Learning – IDEAL 2021* (H. Yin et al., ed.), (Cham), pp. 107–118, Springer International Publishing, 2021.

[2] R. Feldhans and B. Hammer, "Drift Explanation with Difference Accentuation (upcoming)," 2022.

[3] A. Artelt, F. Hinder, V. Vaquet, R. Feldhans, and B. Hammer, "Contrasting Explanations for Understanding and Regularizing Model Adaptations," *Neural Processing Letters*, 2022.

[4] A. Wilke, S. Heindorf, R. Feldhans, B. Hammer, and A.-C. Ngonga Ngomo, "Explaining Drift in Text Data with Document Embeddings (upcoming)," 2022.

EML4U Publications II

[5] M. H. Shaker and E. Hüllermeier, "Ensemble-based Uncertainty Quantification: Bayesian versus Credal Inference," 2021.

[6] S. Schröder, A. Schulz, P. Kenneweg, R. Feldhans, F. Hinder, and B. Hammer, "The SAME score: Improved cosine based bias score for word embeddings," 2022.

[7] H. M. Zahera, R. Jalota, M. A. Sherif, and A.-C. N. Ngomo, "I-AID: Identifying Actionable Information From Disaster-Related Tweets," *IEEE Access*, vol. 9, pp. 118861–118870, 2021.

[8] H. M. Zahera, D. Vollmers, M. A. Sherif, and A.-C. N. Ngomo, "MultPAX: Keyphrase Extraction using Language Models and Knowledge Graphs," in *ISWC*, Springer, 2022.

EML4U Publications III

[9] A. Bondarenko, M. Wolska, S. Heindorf, L. Blübaum, A.-C. N. Ngomo, B. Stein, P. Braslavski, M. Hagen, and M. Potthast, “CausalQA: A Benchmark for Causal Question Answering,” in *COLING*, pp. 3296–3308, 2022.

Thank you for your attention!

- This work has been supported by the German Federal Ministry of Education and Research (BMBF) within the project EML4U under the grant no 01IS19080
- Website: https://eml4u.github.io
- Software: https://github.com/EML4U
- Questions?