Hybrid Learning Approaches

Pattern Recognition and Text Extraction

Shyamnath Premnadh

Barshana Banerjee

Debanjali Biswas

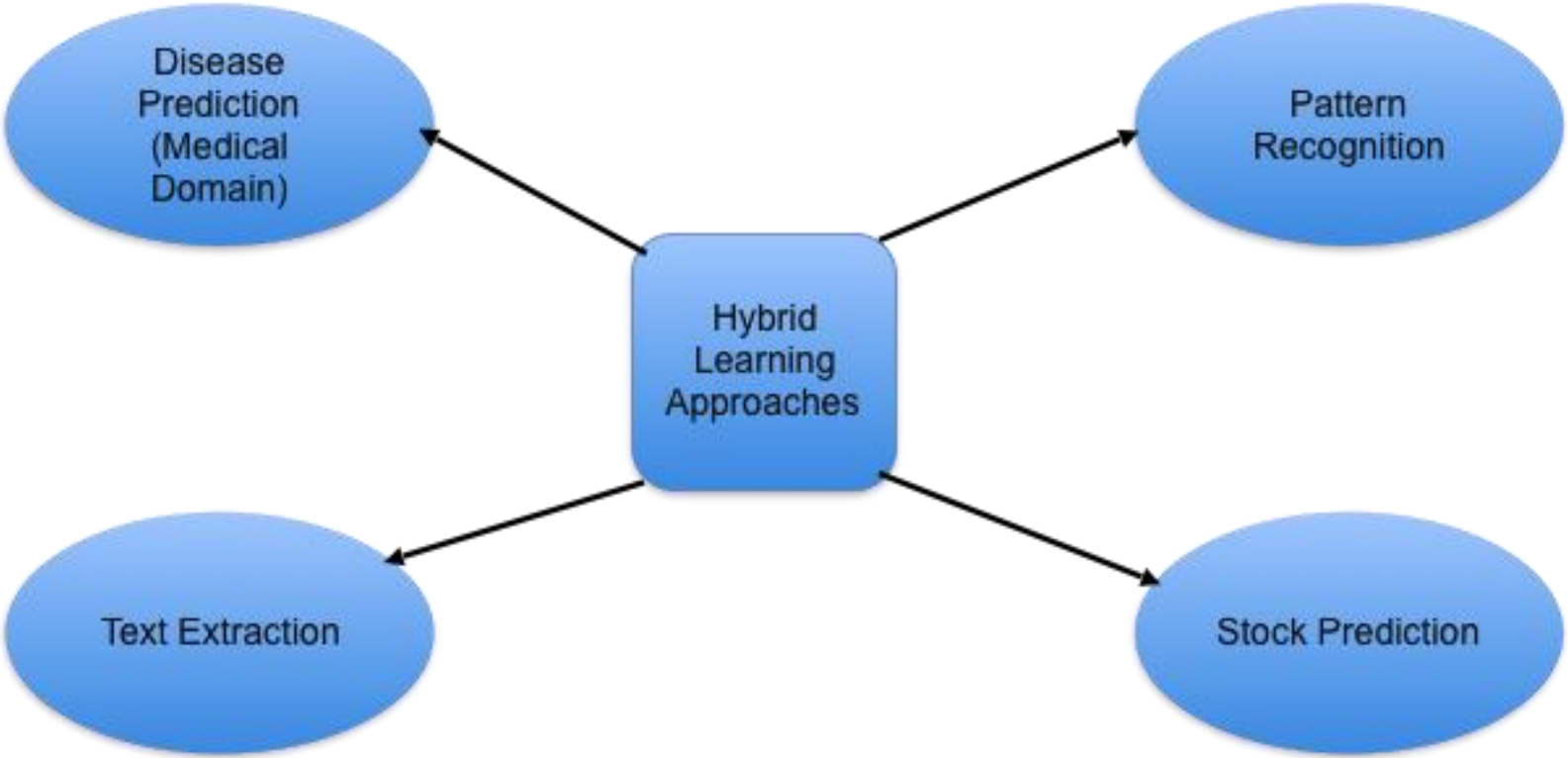
Mst. Mahfuja Akter

Supervisor: Maria Maleshkova

Introduction

Hybrid learning algorithm is a combination of learning paradigms that provide advantages that the original paradigms do not possess.

Use Case / Motivation Scenario/Problem



UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Fraunhofer IAIS

Pattern Recognition

Referred from : Wong Lai Ping, A. T. L. Phuan and Xu Jian

A novel hybrid learning scheme for pattern recognition

Proposed Hybrid Model

The proposed method is a ***self-supervised learning algorithm*** which combines:

- unsupervised clustering K-Means Fast Learning Artificial Neural Network (***KFLANN***)
- with a typical supervised Backpropagation learning algorithm (***BP Network***)

Need for Hybridization

- KFLANN is a unsupervised learning algorithm hence unable to perform interpolation of possible outputs.
- For supervised BP learning algorithm, it is expensive to provide complete target value for most of the real life problem domains.
- To minimize the dependency on external teacher, KFLANN can provide the platform to reproduce these exemplars and thus creating a network to train another network.

Method Description

- KFLANN provide consistent and statistically sound clusters which are used as actual targets in the BP Training Process.
- Input Data is fed to the KFLANN for cluster membership assignment.
- After training, a Cluster Assignment List (CAL) is created.
- MLPBP is trained with the same training data but the teacher value is obtained from the CAL.
- Cluster encoding is direct mapping of the cluster index.

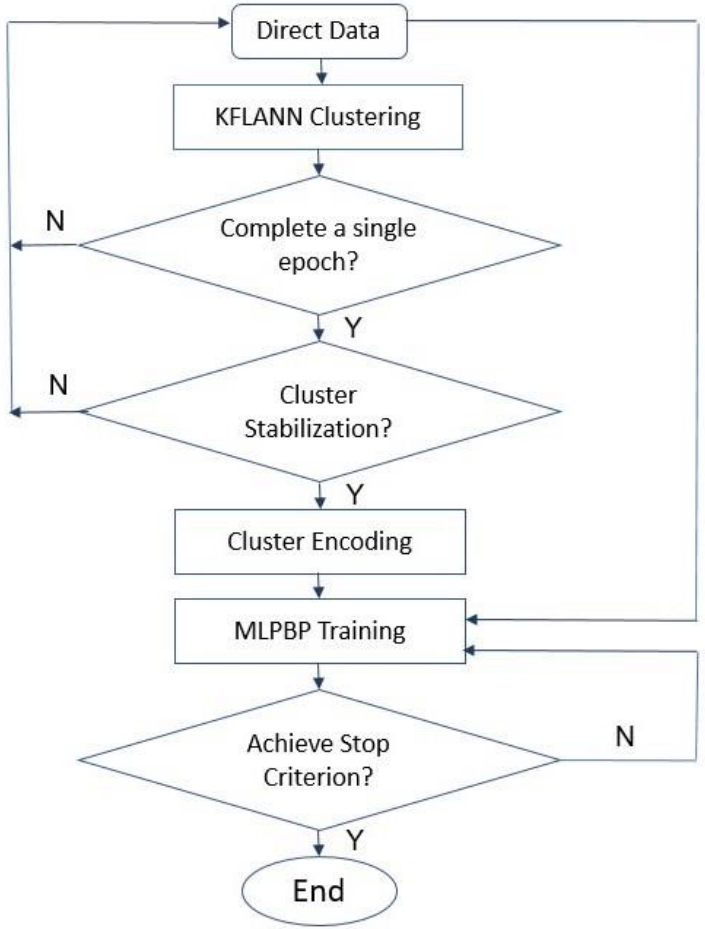


Fig: Flowchart of Hybrid KFLANN and MLPBP Model

Evaluation and Analysis

- Datasets from ***UCI Machine Learning Repository***
- Performance measure : Receiver Operating Characteristics (ROC)
- Accuracy prediction : Area Under Curve (AUC)
- The proposed Hybrid method performs better than KFLANN.
- Although, BP performs slightly better than the Hybrid method, the Hybrid method is self-supervised does not require class information.

Dataset	Hybrid	MLPBP	KFLANN
Iris	96.97	100.00	90.48
Thyroid	89.93	94.67	78.13
Wine	96.67	92.86	90.77
WBC	93.95	94.03	91.77
Dermatology	83.33	80.88	94.44
Zoo	100.00	100.00	100.00
Average	**93.48**	**93.74**	**90.93**

Table: Classification Accuracy

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Fraunhofer IAIS

Text Extraction

Referred from : E. F. A. Silva, F. A. Barros and R. B. C. Prudencio

A Hybrid Machine Learning Approach for Information Extraction

Information Extraction

- Information Extraction (IE) systems => facilitate the information access, by extracting from the documents only the parts that correctly fill in a set of pre-defined output slots (fields).
- Hybrid Approach for IE on semi-structured texts in which an initial extraction performed by a **conventional text classifier(CTS)** is refined through the use of an **HMM.**

Why?

- CTS offers a locally optimal classification without considering the relationships among fragments.

Information Extraction

- **Input :** A semi structured document (in our case a bibliographic reference). A reference is a document with a high variance in its structure.

  Output: Output classes (in our cases it is author, title, affiliation, journal, month, year, editor, place, publisher)

Running Example

Figure represents the proposed approach in the domain of bibliographic references.

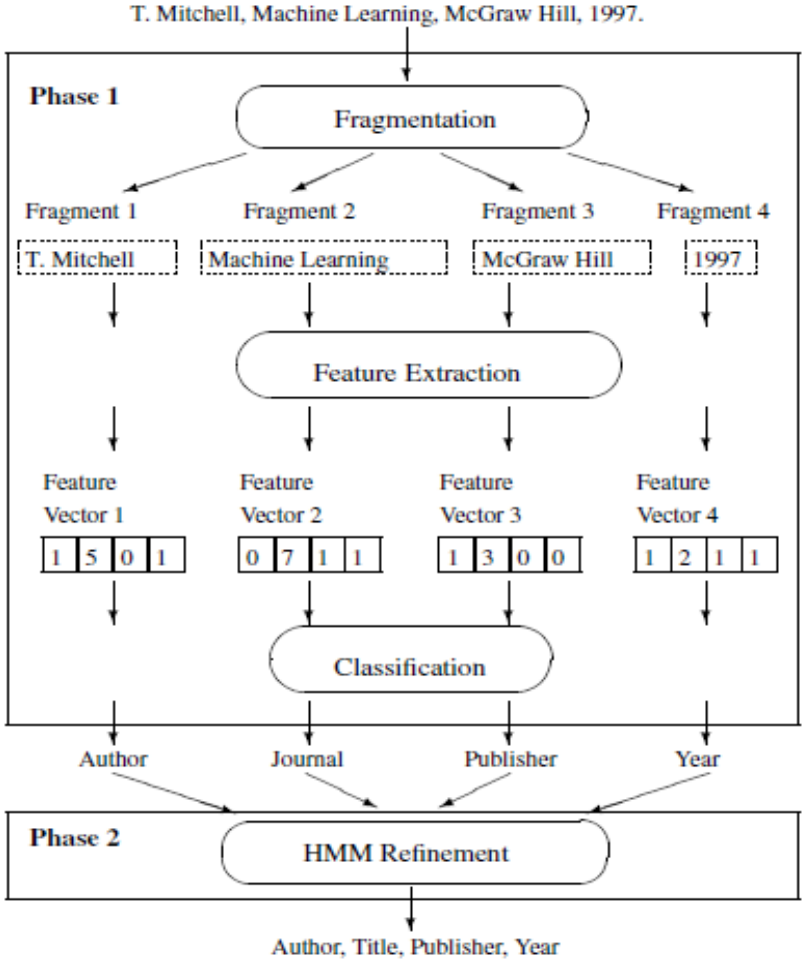


Figure 1. Proposed Approach

Formal Definition of the Approach/ Methods

Phase 1: Extraction using a conventional text classifier

Steps:

1. Fragmentation of the input text – the input text is divided into a set of candidate fragments. This segmentation is performed by a set of heuristics.
2. Feature Extraction – A vector of features for each fragment.
3. Fragment classification - A ML classifier decides which ouput slot will be filled by each input fragment (ML algorithm which is already trained with a corpus of tagged fragments).

Formal Definition of the Approach/ Methods

- Phase 2: Refinement of results using HMM.

The HMM takes as input the whole sequence of class values provided by Phase 1 and returns a refined classification for the given fragments.

1. It has one hidden state corresponding to each slot in the output form.
2. All hidden states were connected to each other

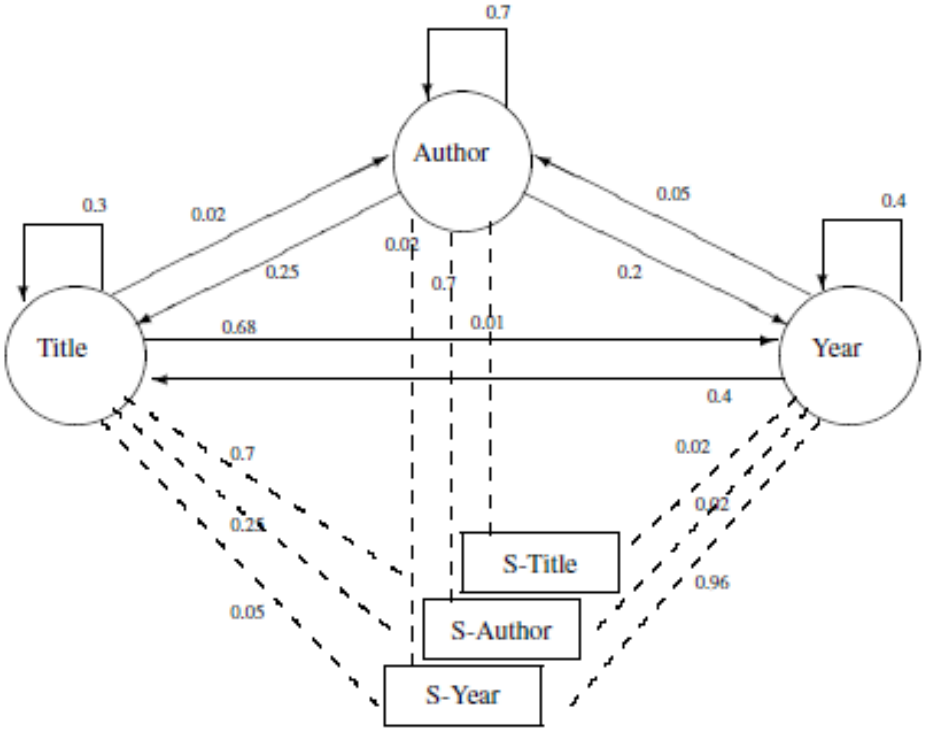


Figure 2. Example of HMM used in Phase 2

Evaluation

- Evaluated using a corpus from a bibliography on computational linguistics which contains 6000 bibliographic references with tags that indicate the class of each text fragment.
- The evaluation measure used was precision, defined as the number of correctly extracted fragments divided by the total number of fragments present in the references.

Table 1. Results obtained in the test corpus

Feature Set	Classifier	Precision without HMM	Precision with HMM	Gain
Manual1	PART	72.17%	76.40%	4.22%
Manual1	Bayes	66.70%	74.72%	8.01%
Manual1	kNN	71.96%	76.28%	4.32%
Manual2	PART	73.48%	77.29%	3.80%
Manual2	Bayes	69.03%	77.27%	8.23%
Manual2	kNN	76.17%	81.16%	4.99%
Automatic	PART	49.91%	72.45%	22.54%
Automatic	Bayes	50.11%	68.25%	18.14%
Automatic	kNN	51.47%	73.57%	22.10%

Conclusion

- Pattern Recognition:
  - The Hybrid method for Pattern Recognition provides an automated means to train the BP network using the KFLANN output as the teacher value.
  - Benefits:
    - Cost effective way learning model
    - KFLANN shares the interpolation capabilities with BP
    - Classification performance is not compromised
- Text Extraction:
  - The novelty of this model was the two techniques have never been combined for an IE system before.
  - The results showed that the use of HMM compensated the relatively low performance of less adequate classifiers and feature sets.
  - Makes it possible to extend to other models.

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Fraunhofer IAIS

Thank You!