Improve Record Linkage Using Active Learning Techniques

CHONG FENG U4943054

SUPERVISED BY DR. QING WANG

DR. DINUSHA VATSALAN

Outline

1. Introduction
2. Motivation
3. Research Problem
4. Methodology
5. Experiments and Evaluation
6. Conclusion and Future Work

Introduction

- **Record linkage** is the process of identifying and matching records that represent the same real-world entity in a database.

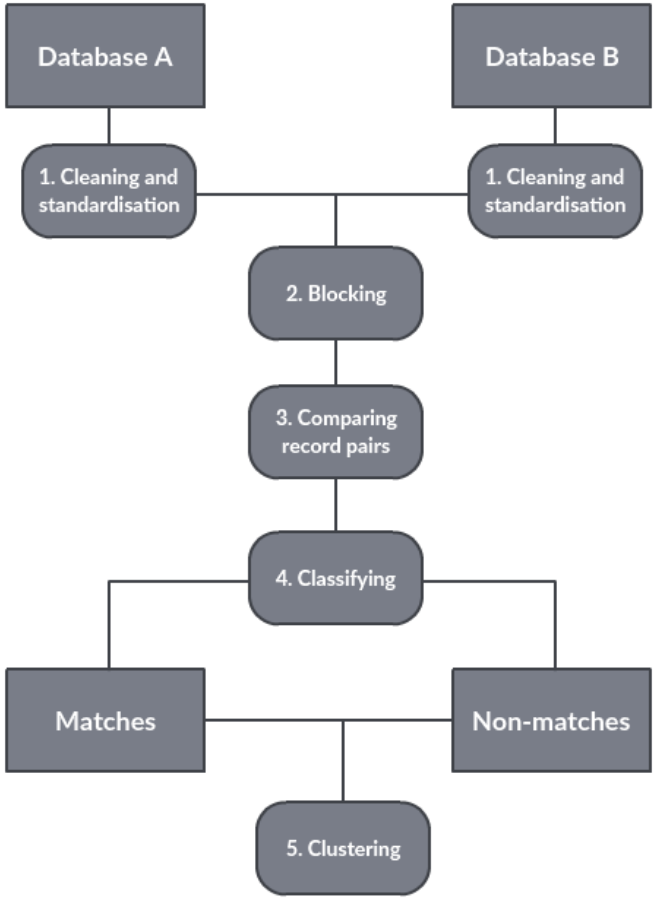
aid	name	affiliation	email
1	Q. Wang		qw@gmail.com
2	Qing Wang	Curtin University	
3	Qing Wang	University of Otago	qw@gmail.com
4	Qing Wang	ANU	qwang@anu.edu.au
5	Qingqin Wang	Curtin University	
6	Wang, Q.		qing@gmail.com
7	Q. Q. Wang	University of Otago	
8	Wang, Qing	University of Otago	
9	Q. Q. Wang	ANU	qw@anu.edu.au

Record Linkage Result

eid	aids
e_1	$\langle 1, 2, 3, 4, 5 \rangle$
e_2	$\langle 6, 7, 8, 9 \rangle$

Example taken from "A Clustering-based Framework for Incrementally Repairing Entity Resolution", by Qing Wang, et al.

Introduction



- Record linkage is based on **similarity weight vectors**, which represent how similar two records are.

DB	Block ID		ID	First Name	Last Name	Suburb	Postcode
A	h400r420	r_1	5620	James	Hawley	Raleigh	27609
	h651m621	r_2	6725	Margaret	Hornback	Mooresville	28117

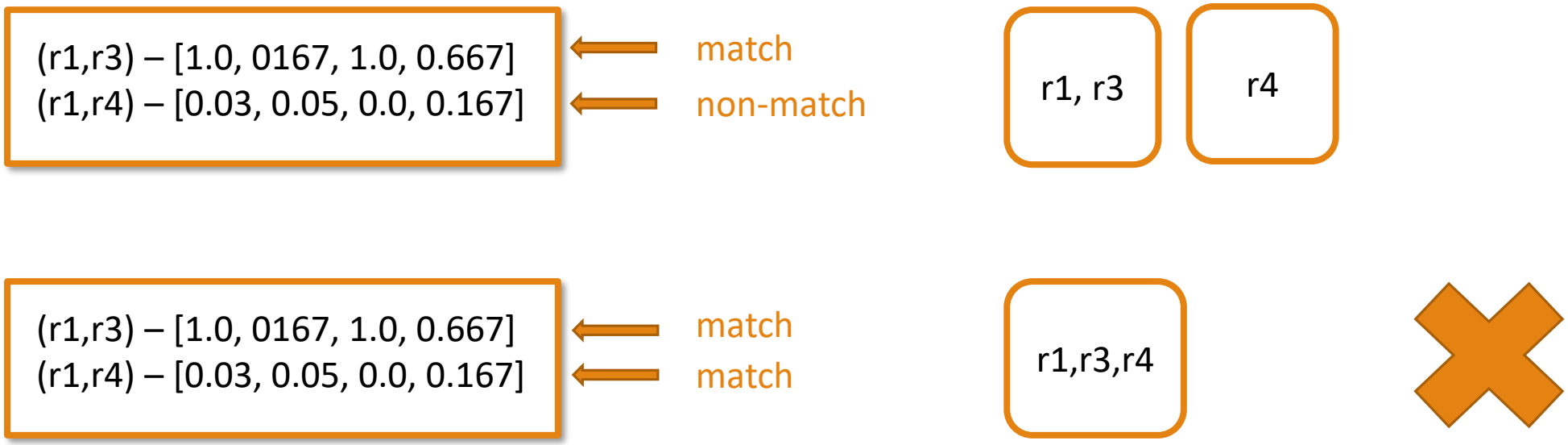
DB	Block ID		ID	First Name	Last Name	Suburb	Postcode
B	h400r420	r_3	7152	James	Hill	Raleigh	27607
		r_4	9845	Bryan	Hill	Raleigh	27601

(r1,r3) – [1.0, 0167, 1.0, 0.667]
(r1,r4) – [0.03, 0.05, 0.0, 0.167]

Similarity weight vectors

Introduction

- In the classification process, weight vectors are classified as either **match** or **non-match** based on their similarity scores.
- A clustering algorithm is used to group records into entity clusters based on their classified weight vectors.



Motivation

- **Errors** may exist in the clustering results, and these errors are normally undetectable in the linkage process.
- These errors might have been introduced into the system from the **classification** process.
- A following question arises:

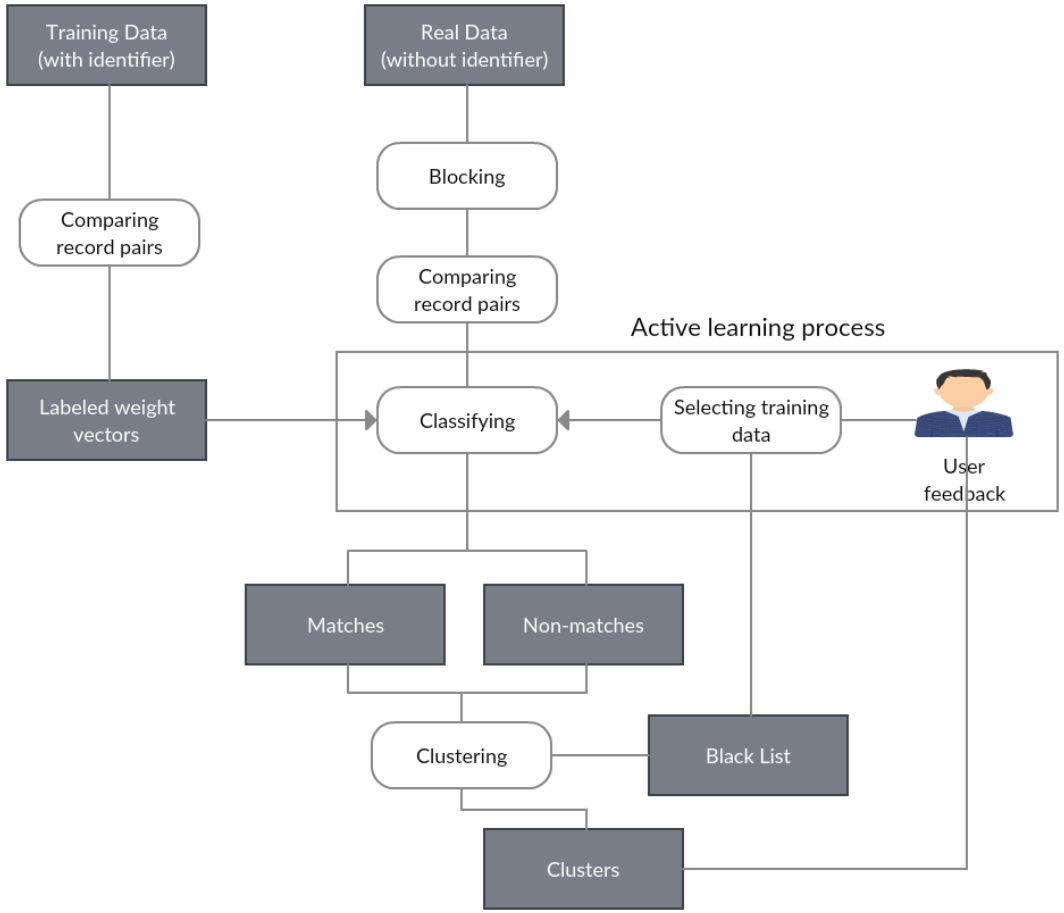
Can we leverage the errors detected by users to improve the classification results as well as the whole record linkage model?

Research Problem

- **Active learning** is a subfield in machine learning and artificial intelligence.
- The idea of active learning is to select the most **informative** training data **iteratively** based on the knowledge it gained from each learning iteration.

- Therefore, the research problem of this project is to find the most informative weight vectors to **refine the classification model** based on user-detected clustering errors.
  - The record linkage model should be able to handle clustering errors, locate the errors in the classification results and repair the errors.
  - The record linkage model should be able to refine the classification model as well as improve the clustering results iteratively using active learning techniques.

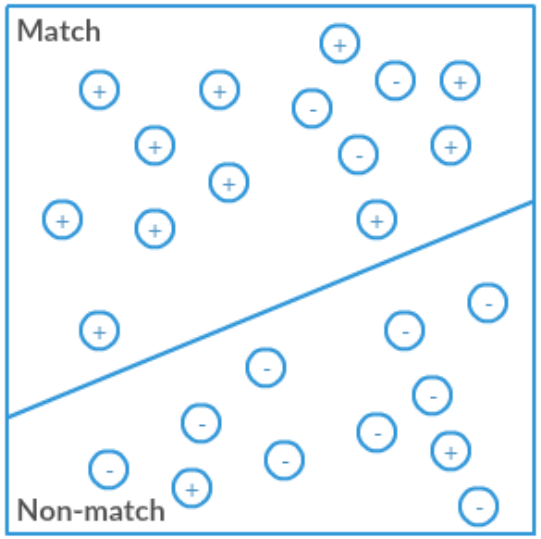
Methodology



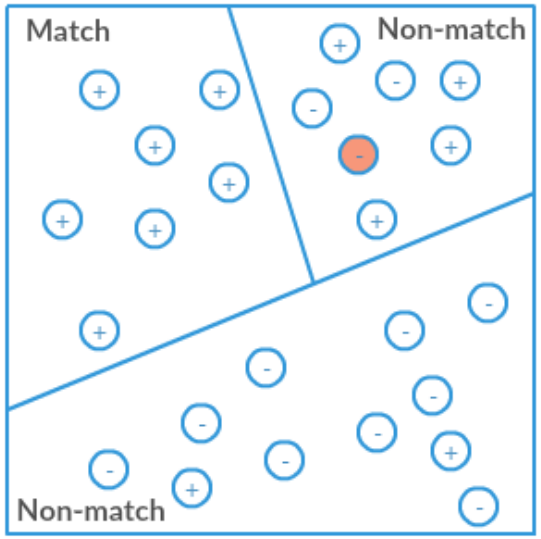
- **An active learning process** is added to the record linkage model, which consists of:
  - user feedback
  - training data selection process
  - reclassification process
- This record linkage model also keeps a **weight vector black list**.
  - Outliers and unrepairable errors will be added to the black list.
  - The black list is updated by the active learning process.
  - Weight vectors stored in the black list will be manually labelled.

Methodology

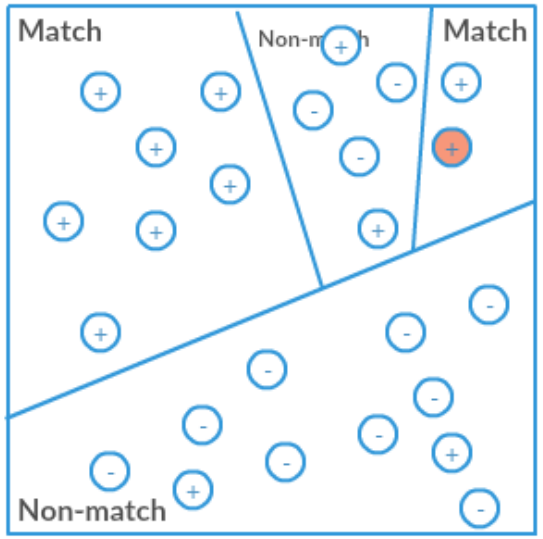
1. Initial classification



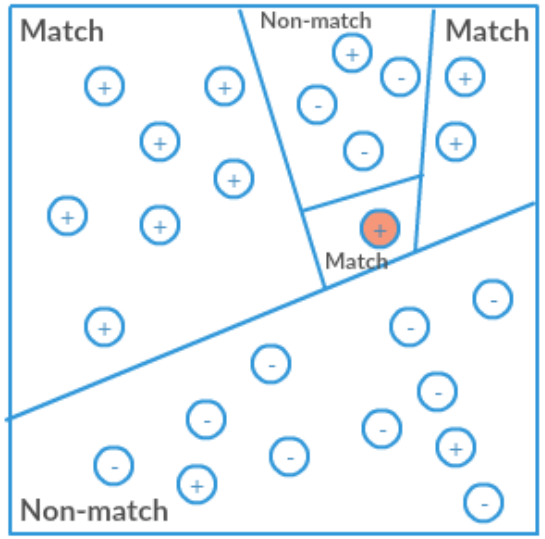
2. First active learning iteration



3. Second active learning iteration



4. After third active learning iteration



Methodology

Algorithm 2 Active learning algorithm

Input:
A list of weight vector spaces: $\mathbf{V}$
A list of training vector spaces: $\mathbf{T}$
A black list of weight vectors: B
An oracle: $Oracle$
A classifier: $Classifier$
An erroneous weight vector: e
Output:
A set of newly classified weight vectors: C

$b = 0$
$C = \{\}$
for vector space $v_i \in \mathbf{V}$ **do**
  if $e \in v_i$ **then**
    $S = \text{MixedSelect}(v_i, e)$
    if $S == \emptyset$ **then**
      B.append(e)
      v_i.pop(e)
    else
      $b = b + |S|$
      S'^m,S'^n= $Oracle$.label(S)
      $S'^m = S'^m \cup \mathbf{T}_i^m$, $S'^n = S'^n \cup \mathbf{T}_i^n$
      $Classifier$.train(S'^m,S'^n)
      W^m, $W^n = Classifier$.classify(v_i)
      if $|W^m| = 0$ or $|W^n| = 0$ **then**
        B.append(e)
        v_i.pop(e)
      else
        $\mathbf{V}$.pop(i), $\mathbf{V}$.append(W^m),$\mathbf{V}$.append(W^n)
        $\mathbf{T}$.pop(i), $\mathbf{T}$.append(S'^m),$\mathbf{T}$.append(S'^n)
        $C \cup W^m$, $C \cup W^n$
      end if
    end if
  end if
end for
return C

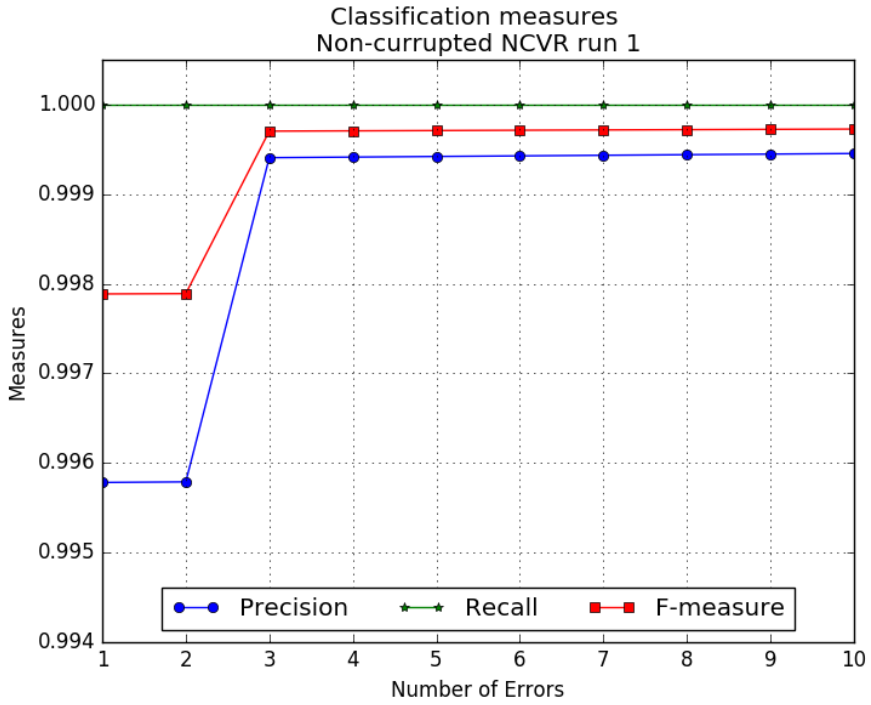
Experiments and Evaluation

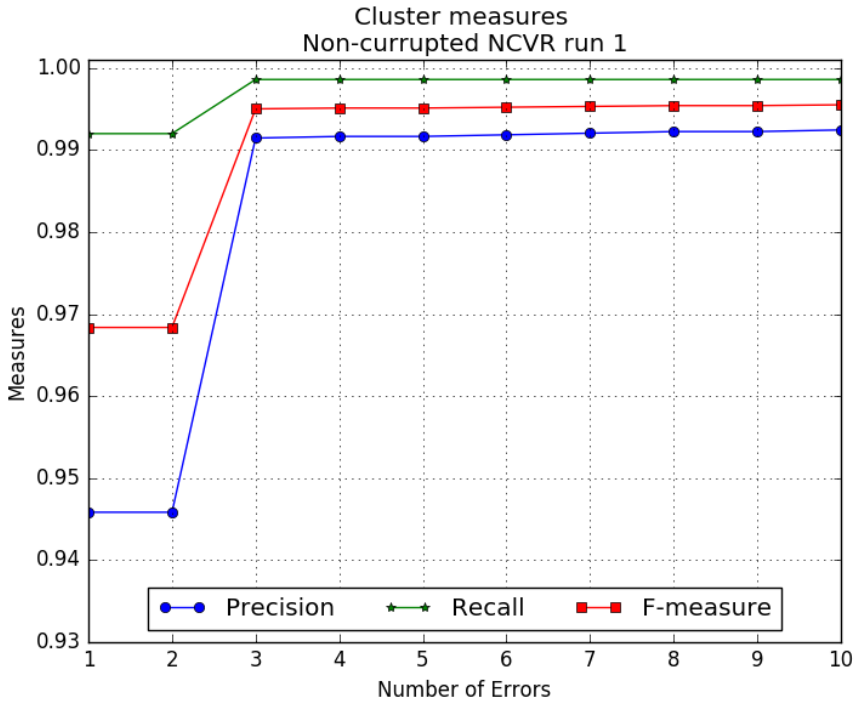
- Datasets:

Dataset	Number of records	Number of weight vectors	Number of clusters	Pair completeness
NCVR Non-corrupted	350,766	100,000	5,000	1
NCVR corrupted	314,663	100,000	5,000	0.5346
CORA	1,878	1,764,381	120	N/A

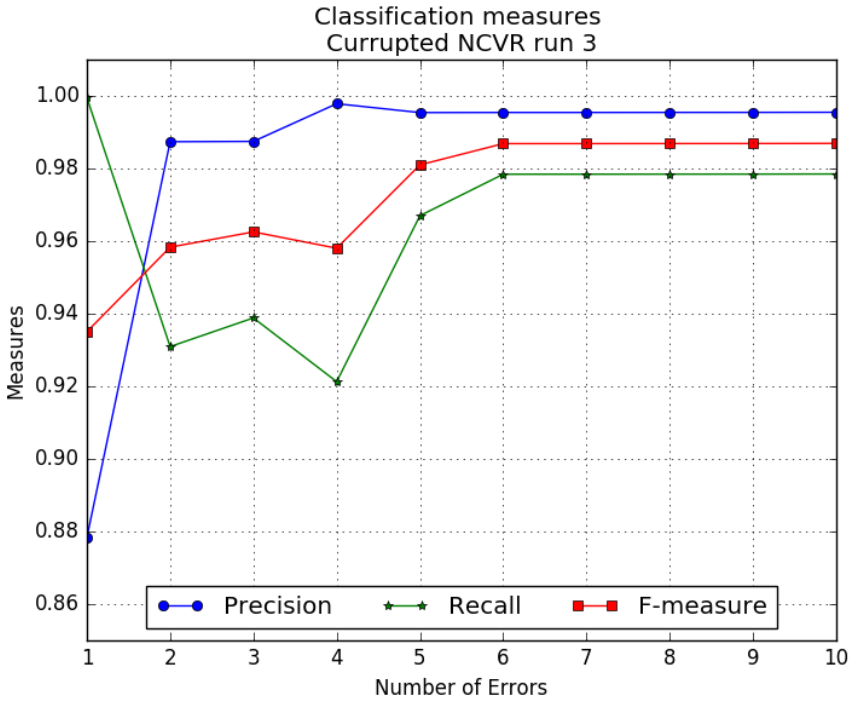
- Measures:
  - Recall, Precision and F-measure

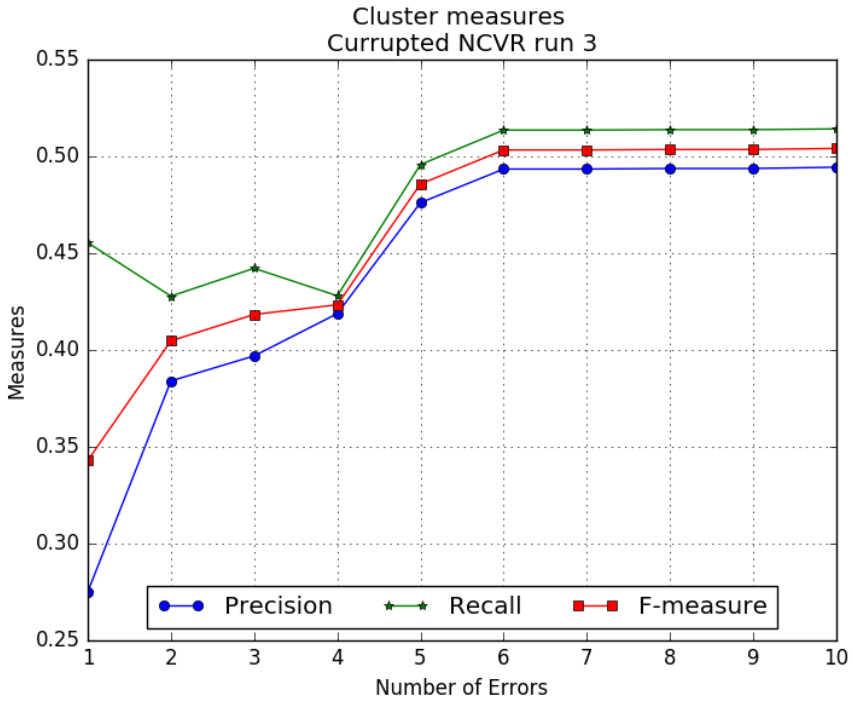
Experiments and Evaluation



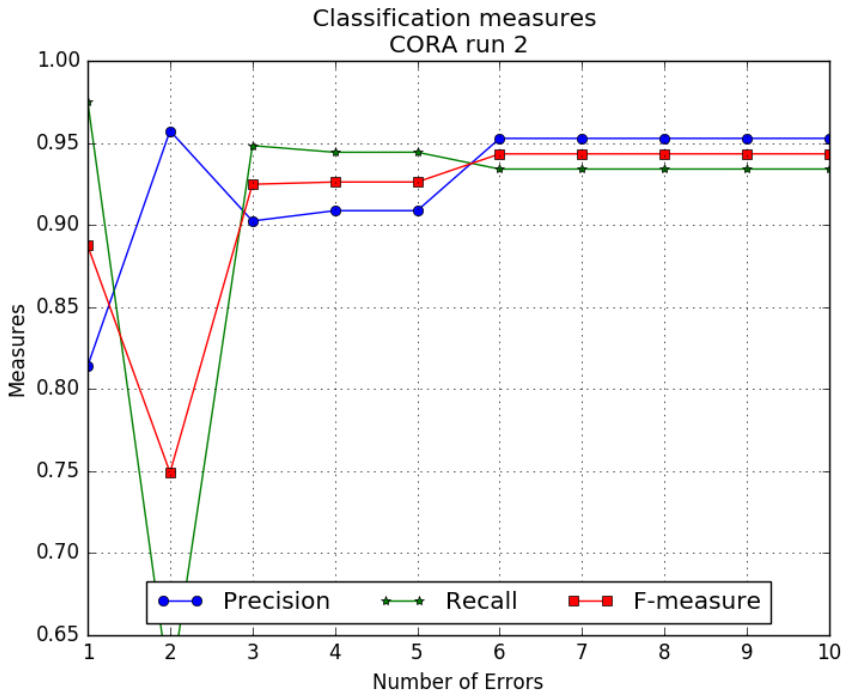


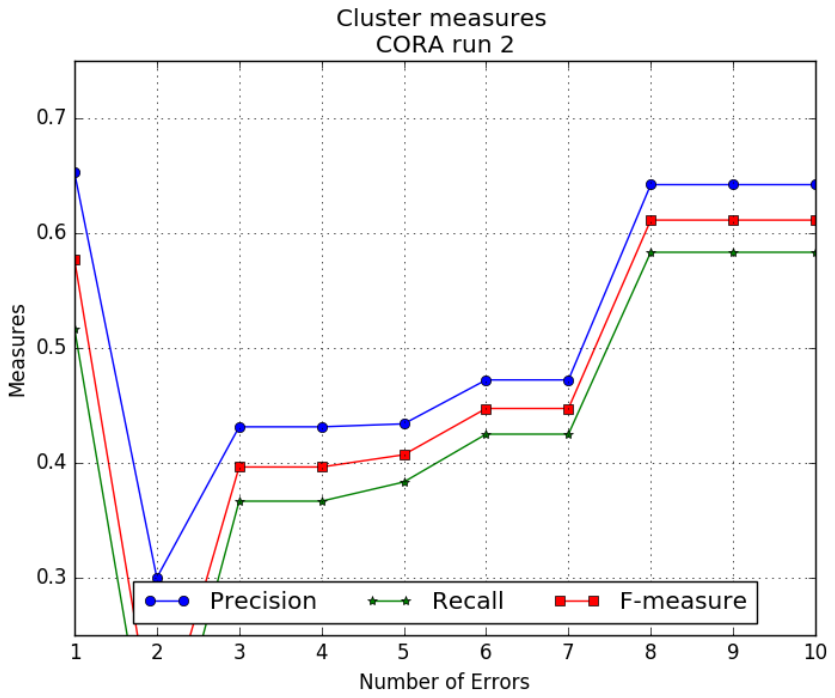
Experiments and Evaluation





Experiments and Evaluation





Conclusion and Future Work

- The experiments proved that the proposed record linkage model can effectively improve classification results and clustering results.
- However, the improvement is also **limited** by other factors.

- Future works:
  - Can we integrate more effective blocking strategies with the record linkage model?
  - How does the selection of errors affect the quality of our models? Can we arrange the order of errors so the weight vector space is divided in the most appropriate way?
  - How to optimize the classifier configuration to achieve the best result?

Thank you!

CHONG FENG U4943054@ANU.EDU.AU