The Australian National University **School of Computing, CECS**

COMP3430/COMP8430 – Data Wrangling – 2022

Lab 6: Evaluation of Record Linkage Week 9

Overview and Objectives

In this lab we are going to build the final step of the record linkage system that we have been working on in labs 3 to 5. Today we look at different evaluation metrics, as discussed in lecture 19, for record linkage, and ask you to implement functions to calculate these measures. Make sure you view lecture 19 before coming into the lab session.

Lab Questions

As in previous labs, please begin by having a look at the overall framework and how each of the Python modules you have created fits together. Today we will be working on `evaluation.py` (available in week 6 in the archive: `comp3430_comp8430-reclink-lab-3-6.zip`). As before, we provide you with some simple code implementations to get started.

We have given you the code to calculate the confusion matrix. The confusion matrix is a table that we use to calculate the performance of the classification technique (or "classifier") on a data set for which the ground truth is known. The confusion matrix itself is relatively simple to understand.

	Predicted True Matches	Predicted True Non-matches
True Matches	True Positives (TP)	False Negatives (FN)
True Non-matches	False Positives (FP)	True Negatives (TN)

As shown in the table above, based on the actual and predicted numbers of matches and non-matches we can divide the linkage result of record pairs into four classes:

- *True positives (TP)*: These are the record pairs which were predicted to be matches, and they are true matches.
- *True negatives (TN)*: These are the record pairs which were predicted to be non-matches, and they are true non-matches.
- *False positives (FP)*: These are the record pairs which were predicted to be matches, but they are true non-matches. Also known as a *Type I error.*
- *False negatives (FN)*: These are the record pairs which were predicted to be non-matches, but they are true matches. Also known as a *Type II error.*

Based on the numbers of record pairs in each of these four classes we can calculate the performance of our record linkage system using different evaluation measures.

Now, manually calculate the evaluation outcomes for the following two confusion matrices:

	Predicted True Matches	Predicted True Non-matches
True Matches	1,000	400
True Non-matches	600	8,000

	Predicted True Matches	Predicted True Non-matches
True Matches	1,200	200
True Non-matches	800	7,800

Apply the following evaluation measures from both above confusion matrices:

- Accuracy
- Precision
- Recall

Which one of the above is the better record linkage outcome? Discuss why.

Moving on to the Python programs, open the module `evaluation.py` in a text editor and have a look at the functions `accuracy()` and `reduction_ratio()` (which is an evaluation metric for the blocking step) which we have provided, to see what the inputs and outputs for these functions are.

Then experiment with these two metrics on the smaller data sets with some of the blocking, comparison, and classification functions you have written previously. Once you are comfortable with how these two metrics are being calculated, then look at implementing the following measures:

1. Precision and recall, which are additional evaluation metrics for record linkage quality.
2. Pairs completeness and pairs quality, which are evaluation metrics for blocking.

All of these were described in lecture 19.

Once you have implemented these measures, please experiment with the smaller data sets provided, and the functions you have implemented in the previous labs. Think about how to test that you have correctly implemented these measures.

As usual, once you have finished your implementation, please experiment with the different data sets provided, and the functions you have been implementing in the previous labs.

Note that the focus of this lab is not on the implementation only, but also about your understanding of how different metrics can be used to evaluate the performance of a record linkage system. This will be important for the upcoming Assignment 3 in this course.

Based on your experiments, some questions you may wish to think about include:

- Are there any measures that are not useful, either because they are always extremely high, or low, or difficult to calculate, etc?
- What is the impact of the corruption level of the data sets on the linkage results, both in the blocking and the final results? Does this vary depending on which functions you use for the blocking, comparison, and classification steps?
- What effect do the different blocking techniques have on the final record linkage results? What does this tell you about when and how to use blocking?