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

COMP3430/COMP8430 – Data Wrangling – 2024

Lab 6: Evaluation of End to End 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.

In the latter part of this lab we will be experimenting and testing how all the different parameters and choices can affect the outcomes of a record linkage project.

Lab Questions: Part 1

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 under Lab 3 - 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 on both the 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.

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. 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 effects do the different blocking techniques have on the final record linkage results? What does this tell you about when and how to use blocking?

Lab Questions: Part 2

Note that the focus of this part of today's lab is about your understanding of how different blocking, comparison, and classification techniques can be used to build a complete record linkage system; and how to select the best performing combination of different techniques based on the obtained linkage result. This will be important for the upcoming Assignment 3 in the Data Wrangling course.

Please first complete any outstanding task or implementations from the previous labs and part 1 of this lab. Once you are done, please download from Wattle the `comp3430_comp8430-rl-additional-datasets.zip` archive (under Lab 6 - Week 9) that contains new data sets for you to experiment with. Then please run the record linkage program on the data sets of different sizes and quality levels (clean to very dirty). Ideally modify your main record linkage program in such a way that it runs linkage on all provided data sets in one go.

Experiment with different function choices in each of the different components (blocking, comparison, and classification). Try different parameter settings for thresholds, different attribute choices for blocking and comparison, different weightings for the classification step, and so on. Some questions you may wish to consider include:

- For each of the different evaluation metrics, which choices produce the best results (or the best results you can find)?
- Are these still the best choices for the different data sets with different sizes and different data quality (corruption) levels?
- Do some of these choices trade-off one evaluation metric against another (i.e. they produce a good result for one evaluation measure but are poor for some others)?
- How significantly does blocking improve the performance (in terms of time)? Does this become more important as the data sets get larger?
- Are there some parameter settings or functions that are worse than others for all data sets and on all evaluation metrics?
- Can you spot any patterns in the results? Are there any functions that seem to work well on different data types? Do certain parameters seem to require a particular range in order to achieve reasonable results (e.g. the similarity thresholds)? Could you use these patterns to justify your choice of functions and parameter settings in the future?

Please note that for some parameter settings and function choices, the program may be very slow, especially on the larger data sets. Please terminate a run early rather than spending the whole lab waiting for it to complete (for some choices it may not finish at all).

In addition there are some other things you may wish to use this lab for such as:

- Write the output (the record id pairs of predicted true matches) of each parameter setting and function choices into a file. You can use the Python program `saveLinkResult.py` which is available under Lab 6 - Week 9 to write the linkage output into a file. Once you downloaded this Python file have a look at the function `save_linkage_set()` which we have provided, to see what the inputs and outputs are of this function. Then call this function from the main Python program `recordLinkage.py` to write the result into a file. Make sure you import `saveLinkResult.py` within `recordLinkage.py` before you call the function `save_linkage_set()`.
- If there are any pieces of the code you did not fully understand or complete, please have another look at them today and ask for assistance from your tutor if you would like further explanations.
- If there were any of the extension exercises from labs 3 to 6 that you partially implemented, then this lab is another opportunity to complete them.
- If you would like to improve some of the components or measurements such as timing or the information printed out, please feel free to do so. You are welcome to customise the program as much as you desire.

Also note that you will have to make use of this program for the upcoming Assignment 3, so please keep this in mind while you are experimenting and make sure you are comfortable with the entire program and know how everything works.