Digital Health Innovation for Consumers, Clinicians, Connectivity and Community
A. Georgiou et al. (Eds.)

doi:10.3233/978-1-61499-666-8-113

An Ensemble Approach for Record Matching in Data Linkage

Simon K. POON[a], Josiah POON[a], Mary K. LAM[b], Qinglan YIN[a], Daniel M-Y. SZE[c], Justin C.Y. WU[d], Vincent C.T. MOK[d], Jessica Y.L. CHING[d], Kam-Leung CHAN[d], William H.N. CHEUNG and Alexander Y. LAU[d]
[a] *School of Information Technologies, The University of Sydney, Australia*
[b] *Faculty of Health Sciences, The University of Sydney, Lidcombe, Australia*
[c] *School of Health and Biomedical Sciences, RMIT University, Australia*
[d] *The Hong Kong Institute of Integrative Medicine, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong*

Abstract. Objectives: To develop and test an optimal ensemble configuration of two complementary probabilistic data matching techniques namely Fellegi-Sunter (FS) and Jaro-Wrinkler (JW) with the goal of improving record matching accuracy. **Methods**: Experiments and comparative analyses were carried out to compare matching performance amongst the ensemble configurations combining FS and JW against the two techniques independently. **Results**: Our results show that an improvement can be achieved when FS technique is applied to the remaining unsure and unmatched records after the JW technique has been applied. **Discussion**: Whilst all data matching techniques rely on the quality of a diverse set of demographic data, FS technique focuses on the aggregating matching accuracy from a number of useful variables and JW looks closer into matching the data content (spelling in this case) of each field. Hence, these two techniques are shown to be complementary. In addition, the sequence of applying these two techniques is critical. **Conclusion**: We have demonstrated a useful ensemble approach that has potential to improve data matching accuracy, particularly when the number of demographic variables is limited. This ensemble technique is particularly useful when there are multiple acceptable spellings in the fields, such as names and addresses.

Keywords. Data Linkage, probabilistic data matching, Fellegi-Sunter, Jaro-Wrinkler

Introduction

In an era where large amounts of patient related health data exist in different data sources, the probabilistic record linkage technique is commonly used to match and link these diverse datasets for patient record analyses. As many of these datasets were collected independently, often at different times and likely for different purposes, the goal of using record linkage techniques is to determine whether the clinical records in different datasets belong to the same patient. As record linkage techniques rely on demographic data to connect records from multiple sources, there are two ways of improving the matching accuracy. One is to identify a number of useful demographic information and the other is to improve the matching techniques for each field. In this project, we attempt to investigate an ensemble approach by integrating two data matching techniques, where the Fellegi-Sunter (FS) technique for combining different demographic information for

probabilistic matching is combined with the Jaro-Wrinkler (JW) technique that accommodates spelling variations in the demographic data. This project is approved by the Institutional Review Board (Joint CUHK-NTEC Clinical Research Ethics Committee) and complies with regulations on data management.

1. FS and JW Data Linkage Methods

The FS approach[1] compares the field values in two records from two different sources. Two sets of probabilities are first determined. M-probability (m_k) is the probability that each field agrees given that the pair of records is matched. U-probability (u_k) is the probability that the given fields agree given that the pair of records is not matched. After that, it calculates a score for each corresponding field of the two records. A final score is then derived based on the aggregation of all scores. The overall formula is shown below:

$$score = \sum_{k=1}^{n} \left(\log\left(\frac{m_k}{u_k}\right)^{\gamma_k^j} \log\left(\frac{1-m_k}{1-u_k}\right)^{1-\gamma_k^j} \right) \text{ for k}^{\text{th}} \text{ field in the j}^{\text{th}} \text{ record pair}$$

Where n = number of identifiers per record
γ_k^j = observed agreement/disagreement value (1 = agree, 0 = disagree)
m_k = estimated identifier agreement rate among links
u_k = estimated identifier agreement rate among non-links

In this technique, there are three decision options: matched, un-matched and unsure. Decision is made based on pre-defined cut-off values. When the pair of records is above the matched threshold, it is labelled "matched". When the pair is below the predefined threshold for unmatched, they are labelled as "un-matched". A pair of records with a value between the thresholds are considered as "unsure". Further steps, such as manual review, is needed for "unsure" pairs to determine if there are additional possible matches.

The JW technique[2, 3] focuses on string comparison between two fields. A score of similarity (x) is calculated based on four pieces of information, including the number of common characters between two strings (c), the length of the two strings (m and n) and half of the number of pairs of common characters that are out of order (t). The score is normalised in the range of 0 to 1, where 0 is dissimilar and 1 is an exact match. This method is especially appropriate for short strings comparison.

$$x = \begin{cases} 0 & if\ m = 0 \\ \frac{1}{3}\left(\frac{c}{m} + \frac{c}{n} + \frac{c-t}{c}\right) & otherwise \end{cases}$$

There are strengths and weaknesses to both approaches. Ensemble learning[4, 5] is an approach to harness multiple models rather than the constituent models on its own to achieve a better computation performance. Ensemble theory considers each solution produced by one model as only providing one view (hypothesis) of a dataset. This view is often incomplete and biased towards the original model, hence insufficient to provide an adequate solution for the learning task. In view of this deficiency, different views can be generated using other models on the same dataset. Hence, multiple models can be used to generate multiple views. The overall performance of the ensemble will depend

on how complementary the models are. The aim of this work is to test the impact and significance of an ensemble approach by employing different sequences combining results from the two linkage methods (FS and JW).

2. Experiments and Analysis

2.1. Data Sources and Data Processing

A dataset with 4156 patient records was obtained from one teaching hospital in Hong Kong. This dataset contained ten fields of personally identifiable information (Record ID, surname, given name, age at admission, sex, date of birth, exact date of birth or not, race, and district of residence). Six fields were used for this study: surname, given name, age of admission, sex, date of birth (DoB), race and district of residence. A field, Record ID, was generated for this project and was kept for evaluation of matching accuracy. All records were stored in an encrypted Excel file. Prior to the generation of experimental datasets, the dataset was checked for duplications. 390 duplicated patient records were removed giving a total of 3766 records in the final dataset.

2.2. Experiments

Two experiments were designed to test the impact of the sequence of employing the two linkage methods in matching results. The two methods were: FS followed by the application of JW (FS-JW) and JW followed by the application of FS (JW-FS). Figure 1 presents the two experimental methods and how the test datasets were generated for the experiments.

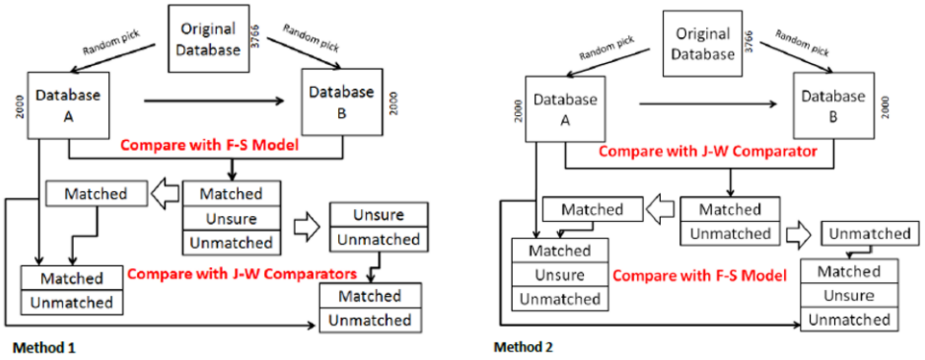

Figure 1. Experimental dataset generation and experimental methods

Datasets for Experiment 1: The random number generation function in Excel was used to generate two experimental datasets of 2000 records (A and B as shown in Figure 1). Both datasets of 2000 records were randomly drawn from the original 3766 records to ensure there is a significant number of overlapping records between datasets A and B.

The goal of the experiment 1 was to test the accuracy for identifying matches and mismatches between the two data subsets of the same size drawn from the same data source of 3766 records, i.e. to accurately match all the overlapping records between the two data subsets.

Datasets of Experiment 2: Using dataset A of 2000 records generated for Experiment 1 as the basis of the generating the second dataset B for matching. Dataset A was manipulated in two ways. These included reducing the dataset by 20% randomly and introducing various errors at a rate of 10% to each column of the remaining records (1600). These errors included typing errors and different ways of spelling names as well as switching to a different category (male, female). The goal of Experiment 2 was to test the accuracy for matching the erroneous data subset of 1600 records to the original dataset of 2000 records, i.e. to accurately match when the fields could not be perfectly matched.

3. Results

The measurements we used for our comparative analyses were sensitivity (recall), specificity, precision, F-measure (which is a harmonic means of the recall and precision) and accuracy. Both experiments were repeated 10 times. The averages of each measurement are presented in Tables 1, 2, 3 and 4.

3.1. Results of Experiment 1

Matching performance between FS and JW models are shown in Table 1. Results show that the F-measure was higher in the FS model than in the JW model, respectively 0.9776 and 0.8920. The difference in matching accuracy was mainly due to the high false positive matches in the JW model. The outcome of the FS model has 4 categories: 1-to-1 matched, 1-to many matched, unsure and unmatched compared to the JW model which has only 3 categories: 1-to-1 matched, 1-to many matched and unmatched. For those records that were in the 1-to-many matched categories (i.e. multiple matches found in dataset B above the threshold to match with the record in dataset A), the record with the highest score was then chosen as the match. For those that were in the unsure categories in FS (i.e. records in-between the thresholds), those records were treated as unmatched.

Comparative results of the ensemble models are shown in Table 2. When considering the application of JW after the execution of FS, the F-measure dropped from 0.9776 to 0.8852. The reduction of performance resulted when JW was applied to the unmatched records generated from FS, which re-introduced a number of false positive matches back into the matched category. When considering the application of FS after the execution of JW, the F-measure improved from 0.8920 to 0.9847. This is due to the ability of the FS model in handling false positive matches from the unmatched category.

Table 1. Results comparing FS versus JW

FS Only		Found		Total
		#Matches	#non-matches	2000
Actual	#Matches	1059.4	0	1059.4
	#non-matches	48.6	892	940.6
Accuracy		0.9757		
Precision		0.9561		
Recall		1.000		
F-measure		0.9776		

JW Only		Found		Total
		#Matches	#non-matches	2000
Actual	#Matches	1059	0.4	1059.4
	#non-matches	265.1	634.5	940.6
Accuracy		0.8718		
Precision		0.8053		
Recall		0.9997		
F-measure		0.8920		

Table 2. Results comparing FS+JW versus JW+FS

FS then JW		Found		Total
		#Matches	#non-matches	2000
Actual	#Matches	1059	0.4	1059.4
	#non-matches	274.2	666.4	940.6
Accuracy		0.8627		
Precision		0.7943		
Recall		0.9996		
F-measure		0.8852		

JW then FS		Found		Total
		#Matches	#non-matches	2000
Actual	#Matches	1059.4	0	1059.4
	#non-matches	33	907.6	940.6
Accuracy		0.9835		
Precision		0.998		
Recall		1.000		
F-measure		0.9847		

Based on the results, we can see the benefits of using the ensemble approach to improve matching performance. However, it is also interesting to see the sequence of the models in the ensemble can affect the overall performance in both directions. In this case, the ensemble of JW-FS has the highest F-measure value.

3.2. Results of Experiment 2

In this experiment, we investigated how each matching model would perform when errors are introduced in the matching process. Results are provided in Tables 3 and 4. Similar patterns emerged from this experiment. When comparing matching performance between FS and JW models, results show that the F-measure was higher in the FS model than in the JW model, respectively 0.9557 and 0.9115. When considering the use of the ensemble approach, the application of FS after the execution of JW achieved the highest F-measure of 0.9602. These results are consistent with findings in the previous experiment.

Table 3. Results comparing FS versus JW

FS Only		Found		Total
		#Matches	#non-matches	2000
Actual	#Matches	1481.8	118.2	1600
	#non-matches	19.2	380.8	400
Accuracy		0.9313		
Precision		0.9872		
Recall		0.9261		
F-measure		0.9557		

JW Only		Found		Total
		#Matches	#non-matches	2000
Actual	#Matches	1381.9	218.1	1600
	#non-matches	50.1	349.9	400
Accuracy		0.8659		
Precision		0.9650		
Recall		0.8637		
F-measure		0.9115		

Table 4. Results comparing FS+JW versus JW+FS

FS then JW		Found		Total
		#Matches	#non-matches	2000
Actual	#Matches	1379.8	220.2	1600
	#non-matches	14	386	400
Accuracy		0.8829		
Precision		0.9899		
Recall		0.8624		
F-measure		0.9218		

JW then FS		Found		Total
		#Matches	#non-matches	2000
Actual	#Matches	1482	118	1600
	#non-matches	4.8	395.2	400
Accuracy		0.9386		
Precision		0.9968		
Recall		0.9263		
F-measure		0.9602		

Results suggest that ensemble with JW-FS combination not only had the highest number of true matches, it also had the highest F-measure in both experiments. JW-FS also seems to be more resilient to existence of errors in data. Interesting, the FS-JW seems to be the least performing approach in data matching.

4. Discussion, Conclusion and Potential Implications

There are a few interesting observations that can be derived from our analyses. First, results suggested that the ensemble approaches do not always improve matching performance, and that the configuration of the models in the ensemble could play a significant role in affecting the results. In our case, the application of the ensemble (FS-JW) was found to be less superior then using FS alone, whereas, significant performance gain could be achieved by apply FS after JW in the ensemble (JW-FS). The non-intuitive insight from this research is the non-trivial complementary effects in relations to the configuration of the ensemble. From the computational perspective, if several models are bundled in the ensemble, a more complex experimental plan would be required to test for optimal configuration.

To conclude, through the use of experiments, we have discovered an ensemble data matching approach that enables the construction of a reliable and convenient platform of

linking records in a probabilistic fashion, using less sensitive demographic data. This scalable method can accommodate new databases for further linkage processes when unique identifiers are not available for matching.

We envisage this ensemble approach to be useful when data linkage is applied to cross country boundaries. For example, if applying our findings to the Asian setting, with the increasing demand on patient privacy and anticipation of larger scale projects in the region, a robust and reliable data linkage method without the need to use patient ID in any form may facilitate big data analytics in different disease models.

One important consideration in record matching is the quality and relevance of the demographic information used for matching. In Australia, emphasis is given to demographic information such as name structure in the format of first name, middle name and last (or family) name, and address. Additionally, the majority of the population reside in standalone houses in relatively low-density population environments. In contrast, an Asian city such as Hong Kong, does not follow the western nomenclature, and reside in high-density apartment blocks. It is not uncommon to have multiple candidates living in the same apartment block with the same surname. This makes addresses less informative in data matching. There are also several methods of spelling (often influenced by dialectic pronunciation) of the name entered in the clinical record database. The results derived from this project have most potential to linking records with Chinese names and for people living in high-density areas.

References

[1] I. Fellegi and A. Sunter, A Theory for Record Linkage, *Journal of the American Statistical Association*, **64** (1969), 1183–1210, doi:10.2307/2286061.

[2] M. A. Jaro, Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, *Journal of the American Statistical Association,* **84** (1989), 414-420.

[3] M. A. Jaro, Probabilistic linkage of large public health data files, *Statistics in medicine,* **14** (1995), 491-498.

[4] D. Opitz and R. Maclin, Popular ensemble methods: An empirical study, *Journal of Artificial Intelligence Research*, **11** (1999), 169–198, doi:10.1613/jair.614.

[5] L. Rokach, (2010), Ensemble-based classifiers, *Artificial Intelligence Review*, **33** (2010), 1–39, doi:10.1007/s10462-009-9124-7.