Journal of Information Technology Management

Print ISSN: 2008-5893

Online ISSN: 2423-5059

The Effect of Transitive Closure on the Calibration of Logistic Regression for Entity Resolution

Yumeng Ye

*Corresponding author, MSC, Department of Information Quality Program, University of Arkansas at Little Rock, Arkansas, USA. E-mail: yxye1@ualr.edu

John R. Talburt

Prof., Department of Information Science, University of Arkansas at Little Rock, Arkansas, USA. E-mail: jrtalburt@ualr.edu

Abstract

This paper describes a series of experiments in using logistic regression machine learning as a method for entity resolution. From these experiments the authors concluded that when a supervised ML algorithm is trained to classify a pair of entity references as linked or not linked pair, the evaluation of the model's performance should take into account the transitive closure of its pairwise linking decisions, not just the pairwise classifications alone. Part of the problem is that the measures of precision and recall as calculated in data mining classification algorithms such as logistic regression is different from applying these measures to entity resolution (ER) results.. As a classifier, logistic regression precision and recall measure the algorithm's pairwise decision performance. When applied to ER, precision and recall measure how accurately the set of input references were partitioned into subsets (clusters) referencing the same entity. When applied to datasets containing more than two references, ER is a two-step process. Step One is to classify pairs of records as linked or not linked. Step Two applies transitive closure to these linked pairs to find the maximally connected subsets (clusters) of equivalent references. The precision and recall of the final ER result will generally be different from the precision and recall measures of the pairwise classifier used to power the ER process. The experiments described in the paper were performed using a well-tested set of synthetic customer data for which the correct linking is known. The best F-measure of precision and recall for the final ER result was obtained by substantially increasing the threshold of the logistic regression pairwise classifier.

Keywords: Entity resolution, Record linking, Machine learning, Logistic regression, Transitive closure.

DOI: 10.22059/jitm.2019.270013.2324

Introduction

Entity Resolution (ER) is the process of determining whether two entity references (references to real world objects) in an information system are referring to the same entity or different entities (Zhou, & Talburt, 2011). Two references referring to the same entity are said to be equivalent references. Each entity has identity attributes, and the values of these attributes help to distinguish one entity from another. For example, the identity attributes for student entities could be first name, last name, student identifier, telephone number, date-of-birth, home address, or other demographic information.

Historically, ER systems have used matching rules to decide if two references are equivalent. The assumption is that the more similar two references are, the more likely they are equivalent. Conversely, the less similar the references are, the less likely they are equivalent. Most ER systems use one of two types of matching strategies. The first matching technique, often referred to as "deterministic" matching uses "if-then" logic with Boolean operators "and/or" to their decisions (Eram, Mohammed, Pillai, & Talburt, 2017). The second strategy, often referred to as "probabilistic" uses the Fellegi-Sunter Model (Fellegi, & Sunter, 1969) of attribute and frequency weights to compute a match score.

More recently, a third strategy is evolving based on machine learning algorithms. Machine learning algorithms such as logistic regression are being applied to the problem record linking and ER (Kobayashi, Eram, & Talburt, 2018). Logistical regression is one of the most natural ML algorithms for entity resolution because its two-way classification of "true" and "false" can easily be interpreted as a "link" or "don't link" decision for pairs of entity references. However, logistic regression belongs to the family of supervised ML algorithms which require training by examples of the desired classifications. In the case of ER, the training is to show the algorithm examples of pairs of references that are equivalent and examples of pairs of reference that are not equivalent.

Even though ER is defined in terms of a pairwise decision process, in its most widely used applications such as master data management (MDM) ER is applied to large numbers of entity references not just a single pair. Extending ER to many references can be somewhat problematic because the decision process applied to pairs of references is not necessarily transitive. For example, if the pairwise decision process is based on similarity (matching), then the fact that A matches B, and B matches C, does not always mean A matches C. Take as a simple example: the deterministic match rule defined by two references should be linked if, and only if, they differ by no more than one character. Following this rule, the reference "MARY" and "MARI" match, and the reference "MARI" and "MERI" match. However, the references "MARY" and "MERI" do not match by this rule.

However, ER is ultimately about equivalence, not matching, and equivalence is transitive. At least equivalence is transitive if one is willing to accept the unique reference assumption. This is the assumption that each entity reference was created to refer to one, and only one, entity (Talburt, & Zhou, 2015). Although poor quality and lack of context may render a reference ambiguous, nevertheless it was created with the intention of referencing one particular entity.

If we think of the references as nodes in a graph, the pairwise decision is whether there should be an edge between two nodes. So, if there is an edge between A and B (a link), and an edge between B and C, then A and C are still connected even though there is not an edge between A and C. If we are confident the links predict equivalence, and accept the unique reference assumption, then A, B, and C are all equivalent.

In short, if we want to know all of the references to the same entity, we must find the transitive closure of the pairwise links. In graph theory, these maximally connected subsets of a graph are called components of the graph. In ER terminology the components are often called “clusters.”

Logistic Regression is a supervised classification machine learning algorithm, which predicts a binary dependent outcome with a set of independent variables provided. It fits the data into a logit function and then predicts the probability of occurrence of an event, which is shown as the logarithm of odds. The logit model can be expressed as the following regression equation:

1) $$Logit[\hat{p}] = ln\left(\frac{\hat{p}}{۱ - \hat{p}}\right) = \hat{\beta}_۰ + \hat{\beta}_۱ X_۱ + \hat{\beta}_۲ X_۲ + \cdots + \hat{\beta}_p X_p$$

Where $\hat{p}$ is referred as the probability equal to 1, which means the event occurs, and $۱ - \hat{p}$ means the probability equal to 0, which the event does not occur. In addition, the error is distributed as standard logistic distribution, and $\widehat{\beta}$ is the regression coefficient of the predictor variable. The expected probability of $\hat{p}$, which the event occurs, can be computed from the regression equation with the values of X are given.

2) $$\widehat{p} = \frac{exp(\widehat{\beta}_۰ + \widehat{\beta}_۱ X_۱ + \widehat{\beta}_۲ X_۲ + \cdots + \widehat{\beta}_p X_p)}{۱ + exp(\widehat{\beta}_۰ + \widehat{\beta}_۱ X_۱ + \widehat{\beta}_۲ X_۲ + \cdots + \widehat{\beta}_p X_p)}$$

Logistic Regression in Entity Resolution

For ER, the logistic regression model must be applied to pairs of references and predict whether the references are equivalent or not equivalent, i.e. should be linked or not linked, respectively. If the truth set is available, a training dataset can be built and then apply logistic regression to train the classification model (Christen, 2014). The approach taken in this paper is to convert a pair references into a vector of similarity scores. Because the data used in the experiment was structured, a pair of references also corresponds to a set of attributes pairs. The attributes selected for this experiment were First Name, Last Name, Address, City, State, Zip, Social Security Number (SSN). Table I show an example of four references from the test dataset.

Table 1. Input Data

RecID	First Name	Last Name	Address	City	State	Zip	SSN
1	Barbie	Chavze	**11r881 Gulf Pointe** Dr	Houston	Texas	**77089**	**100525974**
2	Brabara	Chavez	**2943 N Cottonwood** St	Orangec	California	**92865**	**100525974**
3	Barbaar	Myers-Christian	**3536 N Berlin Ave**	Fresno	California	**93722**	**10525974**
4	Clyde	Myers-Christian	**10237 Casanes Ave**	Downey	California	**90241**	**10534576**

The rows in Table II show the vectors generated by each of the six possible pairs of the four references in Table I. The vector comprises the Normalized Levenshtein Edit Distance (LED) between the values of the corresponding attributes such as First Name, Last Name, and so on. Each LED similarity score is calculated through the number of edits between two strings [8]. The final term of vector is a bit indicating whether the references are equivalent, 1 for equivalent 0 for not equivalent. These vectors are the form of input data used to train the logistic regression model.

Table 2. Vectors of LED Scores for Pairs of References from Table I

RecID	First Name	Last Name	Address	City	State	Zip	SSN	Equivalent
1,2	**0.29**	**0.67**	**0.10**	**0.00**	**0.00**	**0.00**	**1.00**	**1**
1,3	**0.57**	**0.13**	**0.14**	**0.29**	**0.00**	**0.00**	**0.89**	**1**
1,4	**0.17**	**0.13**	**0.14**	**0.14**	**0.00**	**0.00**	**0.44**	**0**
2,3	**0.43**	**0.13**	**0.20**	**0.14**	**1.00**	**0.20**	**0.89**	**1**
2,4	**0.00**	**0.13**	**0.20**	**0.29**	**1.00**	**0.20**	**0.44**	**0**
3,4	**0.00**	**1.00**	**0.29**	**0.00**	**1.00**	**0.20**	**0.50**	**0**

Transitive Closure in Entity Resolution

Finding the maximally connected subsets of linked pairs requires the application of transitive closure (Talburt, Zhou, & Shivaiah, 2009). The final result of ER is viewed in terms of these graph component or clusters representing the sets of equivalent references. Furthermore, precision and recall, the generally accepted measures for ER performance are computed based on the final clusters of equivalent reference, not on the individual pairwise linking decisions.

Experimental Design

The experiment was conducted to measure the accuracy of using a logistic regression model for ER before and after the application of transitive closure. The experimental procedure can be divided into the training part and the testing part. The algorithms described in this paper were performed using Python to train and test the logistic regression model.

The training part includes the following steps:

١. Begin with a set of data which the references are already clustered into groups with a column that shows their group number (truth set T).

٢. Randomly split a portion of truth set as the training set A where $A \in T$.

٣. Create the set P which contains all the pairs of references in the training set A. i.e. if set $A = \{A1, A2, A3\}$, then $P = \{(A1, A2), (A1, A3), (A2, A3)\}$

٤. For each pair in P, generate a vector which the length equal to one plus the number of identity attributes.

٥. The value of each component in the vector is the LED score between the values of the two corresponding references, followed by the order of identity attributes.

٦. The last value in the vector is the Boolean value (0 or 1), which 0 means a non-equivalent pair and 1 means an equivalent pair. For example:

If there are three identity attributes in the truth set T, First Name, Last Name and Date of Birth. The vector will contain four values, which are the similarity score of First Name, Last Name, Date of Birth, and the Boolean equivalence.

٧. After all vectors from P are generated, the vectors are the input of training data for logistic regression model where the last Boolean value is the dependent variable and the rest of the values are independent variables.

After the above seven steps, the logistic regression model is built and trained. To test it, the model is applied to the rest of B references which were not used to build the model in the truth set T, i.e. $B = T - A$. The testing part include the following steps:

١. Create the set Q which contains all the pairs of references in the testing set B. i.e. if set $B = \{B1, B2, B3\}$, then $Q = \{(B1, B2), (B1, B3), (B2, B3)\}$

٢. For each pair in Q, generate a vector which the length equal to the number of identity attributes.

٣. The value of each component in the vector is the LED score between the values of the two corresponding references, followed by the order of identity attributes.

٤. Use the values in the vector as the predictor variables for the logistic regression model, which built in training part.

٥. Predict the matching probability for each pair of references. Then determine the threshold to match equivalent pairs to receive the matching results (0/1).

٦. Compare the predicted matching result with the truth matching result. Then count the number of true positive, true negative, false positive and false negative of the results before transitive closure.

٧. Apply the transitive closure algorithm to cluster all references in the testing set B. Assign an identifier to each reference. The same identifier represents the same cluster. The transitive closure procedures of the experiment are designed listed below:

a) Create a table which contains the Record ID of each reference. Add a new column (Group ID) as the clustering identifier. Set the initial value to be 0.

b) Create an empty hash table which is used for storing clustering information of references.

c) Initialize an integer (group_id) to 1, as a dynamic pointer to the group_id of each reference.

d) Initialize an integer (c1), set to zero as the group_id of the first reference.

e) Initialize an integer (c2), set to zero as the group_id of the second reference.

f) For each pair of records in the testing set B, after each prediction, assign new value of c1 and c2 by the following deterministic logic:

Case 1: If both c1 and c2 equal to zero and the prediction shows two records are matched. Set the value of c1 and c2 to group_id. Then add c1 as a key of the hash table, two records as values.

Case 2: If both c1 and c2 equal to zero but the prediction shows two records are not matched. Set the value of c1 to group_id, set the value of $c2$ to $c1 + ١$, increment group_id by 1. Then add c1 as a key of the hash table, the first record as the value. Add c2 as a key of the hash table, the second record as the value.

Case 3: If c1 does not equal to zero, c2 equals to zero, and the prediction shows two records are matched. Set c2 equal to c1. Then add the second record as a corresponding value to the key c1 in the hash table.

Case 4: If c1 does not equal to zero, c2 equals to zero but the prediction shows two records are not matched. Increment the group_id and c2 by 1. Then add c2 as a key of the hash table, the second record as the value.

Case 5: If both c1 and c2 not equal to zero, and the prediction shows two records are matched, and the value of c2 is greater than c1, move all the values in the key c2 to the key c1, delete the key c2.

Case 6: If both c1 and c2 not equal to zero, and the prediction shows two records are matched, and the value of c2 is smaller than c1, move all the values in the key c1 to the key c2, delete the key c1.

After the transitive closure is completed, a dataset is created which contains the RecID of each reference and their Group ID. Then compare the grouping of the truth set and count the number of true positive, true negative, false positive and false negative of the results after transitive closure.

Measurement of Performance

The generally accepted measure for evaluation both ML classifiers and ER results are precision, recall, and F-measure. Although accuracy would seem like an obvious choice, it does not work well for most ER applications because of the large volume of true negative pairs expected. Precision and recall are calculated using only the true positive (TP), false positive (FP), and false negative (FN) counts. The F-measure is the harmonic mean of precision and recall. The formulas for these measures are shown here.

3) $$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

4) $$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

5) $$F1 = \text{۲} * \frac{Precision * Recall}{Precision + Recall}$$

The precision (also called positive predictive value) shows that if two references are predicted to be equivalent and linked by the process, how likely are the references to equivalent. The recall (also called true positive rate) shows that for all pairs which are equivalent in the truth set, what proportion were actually classified or linked as equivalent by the process. Therefore, F-measure shows the test's accuracy by considering both precision and recall computing the score.

Experiment Description

The experiment used a set of synthetic person data created by the Synthetic Occupancy Generator (SOG) (Talburt, Zhou, & Shivaiah, 2009). By using synthetic data, it is easy to know the truth set of each reference. This dataset has been used in graduate level of ER courses by a couple of teams of students using the traditional rule-based matching approach.

The original synthetic data has three lists containing 271,142 entity references. The synthetic data used for this experiment consists of a subset of the original data with 727 references which belong to 150 entities (clusters). The entities simulate people living in a series of U.S. residence, and for each simulated individual, there is an occupancy history of address changes and in some cases, name changes. In addition, some data quality issues, such as missing values, inconsistent formats, misspellings, were injected into the references. The dataset contains the following identity attributes: First Name, Last Name, Street Address, City Name, State Abbreviation, Zip Code, and Social Security Number (SSN).

For each of 263,901 pairs of references ($pair\_count = ٧٢٧ * (٧٢٧ - ١)/٢$), the LED value was calculated and to be used as either the training set or the testing set. In this experiment, 50% of the data were randomly selected for the training set, and the other 50% of the data was selected to the testing set. After the model is trained, the test phase was done 9 times. In each time, the threshold which determines the probability of linking (matching), was set differently from 0.10 to 0.90.

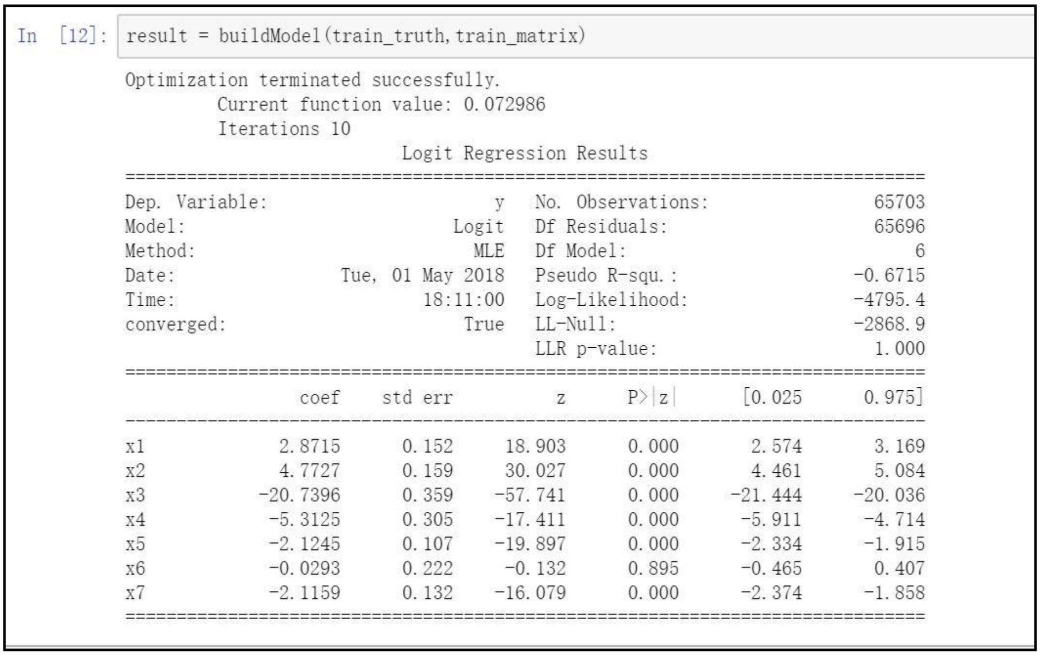
Experiment Results

Table III shows the detailed number of true positive, true negative, false positive, false negative, precision, recall and F-measure before and after the transitive closure.

Table 3. Overview of Test Results

	Before Transitive Closure						After Transitive Closure					
TH	**TP**	**FP**	**FN**	**P**	**R**	**F**	**TP**	**FP**	**FN**	**P**	**R**	**F**
0.1	432	157	81	0.73	0.84	0.78	465	1574	48	0.23	0.91	0.36
0.2	411	91	102	0.82	0.80	0.81	460	586	53	0.44	0.90	0.59
0.3	388	65	125	0.86	0.76	0.80	442	231	71	0.66	0.86	0.75
0.4	376	42	137	0.90	0.73	0.81	443	174	70	0.72	0.86	0.78
0.5	361	29	152	0.93	0.70	0.80	419	112	94	0.79	0.82	0.80
0.6	354	21	159	0.94	0.69	0.80	410	78	103	0.84	0.80	0.82
0.7	337	13	176	0.96	0.66	0.78	397	62	116	0.87	0.77	0.82
0.8	316	9	197	0.97	0.65	0.75	384	56	129	0.87	0.75	0.84
0.9	271	3	242	0.99	0.54	0.69	334	9	179	0.97	0.65	0.78

Figure 1 shows the output of the fitted logistic regression model. The middle section provides some statistical outputs of the results including degree of freedom residuals, pseudo R-squared and log-likelihood etc. On the bottom section of the output, x1 through x7 refer to seven attributes in the model. Then the coefficient, standard deviation, z-score, p-value and 95% confidence interval are shown corresponding to each attribute.



```
In  [12]: result = buildModel(train_truth, train_matrix)

          Optimization terminated successfully.
                   Current function value: 0.072986
                   Iterations 10
                                     Logit Regression Results
          ==============================================================================
          Dep. Variable:                      y   No. Observations:               65703
          Model:                          Logit   Df Residuals:                   65696
          Method:                           MLE   Df Model:                           6
          Date:                Tue, 01 May 2018   Pseudo R-squ.:                -0.6715
          Time:                        18:11:00   Log-Likelihood:               -4795.4
          converged:                       True   LL-Null:                      -2868.9
                                                  LLR p-value:                    1.000
          ==============================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
          ------------------------------------------------------------------------------
          x1             2.8715      0.152     18.903      0.000       2.574       3.169
          x2             4.7727      0.159     30.027      0.000       4.461       5.084
          x3           -20.7396      0.359    -57.741      0.000     -21.444     -20.036
          x4            -5.3125      0.305    -17.411      0.000      -5.911      -4.714
          x5            -2.1245      0.107    -19.897      0.000      -2.334      -1.915
          x6            -0.0293      0.222     -0.132      0.895      -0.465       0.407
          x7            -2.1159      0.132    -16.079      0.000      -2.374      -1.858
          ==============================================================================
```

Figure 1. The output of fitted logistic regression model

Figure 2 visualized the comparison of F-measure before and after transitive closure into the line plot. It shows that before 0.5 threshold, the performance before transitive closure is better than that of the after. At 0.5 threshold, two performances are at the same level. Then when the threshold is greater than 0.5, the performance after transitive closure is better than that of the before.

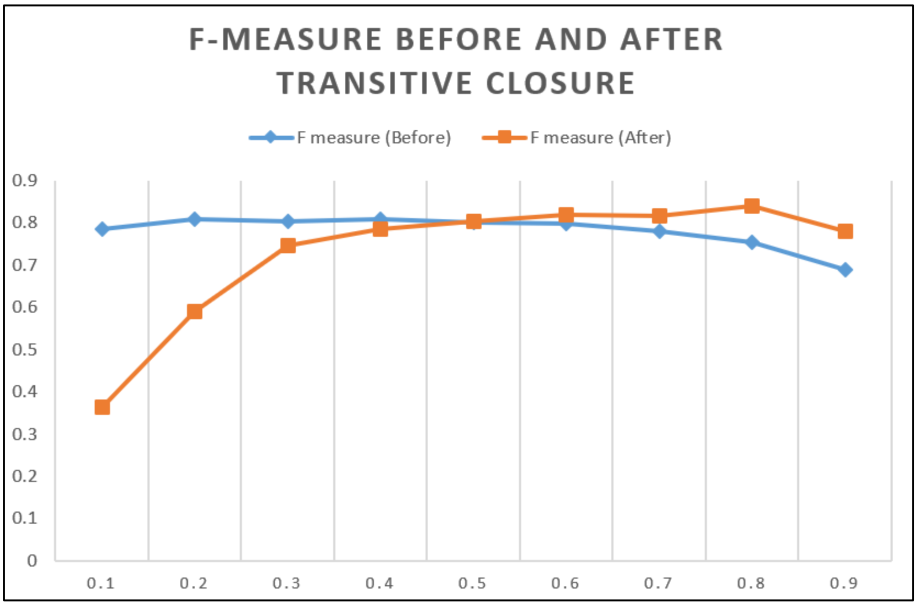


Figure 2. F-measure comparison before and after transitive closure

Summary of Results

This paper described the application of logistic regression machine learning to create an ER process. The truth set of input references are correctly clustered into equivalent references each labeled with a cluster identifier. References with same cluster identifier are known to be equivalent, and references with different cluster identifiers are known to not be equivalent. However, it is important to note the cluster identifier does not convey any information regarding the similarity of the references. Two references can be quite dissimilar but still be in the same cluster because they are referring to the same entity, e.g. patient.

Herein lies the crux of the problem. During training, when the classifier is given two references known to be equivalent, the classifier is looking for some pattern of similarity upon which to recognize the references as equivalent. So in effect, equivalent references are used to train the classifier in similarity or matching. However, after the model is trained and made its pairwise decision, the transitive closure process is then applied to find the clusters of equivalent references. The effect of transitive closure is to make additional links between references. These are inferred links not applied by the pairwise classifier.

Returning to the simple example, if the pairwise classifier links A with B and links B with C, but does not link A with C, the transitive closure process will essentially add the link between A and C by putting A, B, and C into the same cluster of equivalent references. Also the precision and recall measures of these two processes will be different. If A, B, and C are all actually equivalent, then by linking A and B, and linking B and C, and not linking A and C, the pairwise classifier process is evaluated as having made 2 true positive decisions and one false negative decision. On the other hand, in the evaluation of the final ER after transitive closure, A, B, and C are all placed in the same cluster, the pairs A and B, B and C, and A and C are all counted as true positives.

The same effect can be observed in the experimental results shown in Table III. The logistic regression classifier achieves its best F-measure of 0.810 at a threshold of 0.2. Such a low threshold is aggressive in its linking, but necessary in order to recall a significant number of true positive links for relatively dissimilar pairs. However, the transitive closure while recalling 49 additional true positive also adds another 175 false positive links.

On the other hand, the best ER result is achieved when the threshold is raised to around 0.8. At this level the classifier is acting more like a similarity function and linking only very similar pairs. When transitive closure is applied to these links, an additional 68 true positive links are recalled while only increasing the number of false positives by 47. Here the classifier and the transitive closure are working together in a more cohesive fashion. The classifier is focused more on similarity and the transitive closure is chaining through the references to find additional equivalencies.

This raises the question in general for use of supervised machine learning for ER. Are you training the algorithm to find matches (i.e. similarities) or equivalencies? The conundrum is that the starting point, as in these experiments, is usually a training set of equivalent references, not necessarily similar references. At the same time, teaching the algorithm similarities based on traditional matching algorithms does not seem to provide any advantage and begs the question: why not just use traditional matching to start with? The obvious answer is ML does not require the laborious analysis and introspection. At the same time, it is important to understand that when applying these techniques to ER, the thresholds that give the best pairwise classification do not necessarily give the best ER result.

Future Work

The question unanswered by the research presented here is how to adjust the calibration of the ML classifier to yield the best ER result. With the test data and logistic regression classifier used in these experiments it seems the re-calibration is significant from 0.2 to 0.8. Whether this is true of other types of data and other classifiers are questions for additional research.

References

Christen, P. (2014). *Data Matching Concepts and Techniques for Record Linkage. Entity Resolution, and Duplicate Detection*. Berlin: Springer Berlin.

Eram, A., Mohammed, A.G., Pillai, V. & Talburt, J.R. (2017). Comparing the Effectiveness of Deterministic Matching with Probabilistic Matching for Entity Resolution of Student Enrollment Records. *MIT International Conference on Information Quality*, Little Rock, AR, Oct 6-7.

Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.

Kobayashi, F., Eram, A., & Talburt, J. (2018). Comparing the Performance of Logistic Regression Classification to Rule-Based Entity Resolution. *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. DOI: 10.1109/mipr.2018.00033.

Talburt, J. R., Zhou, Y., & Shivaiah, S. Y. (2009). SOG: A synthetic occupancy generator to support entity resolution instruction and research. *2009 International Conference on Information Quality*, Potsdam, Germany, pp. 91-105.

Talburt, J.R. & Zhou, Y. (2015). *Entity information life cycle for Big Data: Master data management and information integration*. Elsevier. Waltham, MA.

Zhang, J., Bheemavaram, R., & Li, W. N. (2009). Transitive Closure of Data Records: Application and Computation. *Data Engineering International Series in Operations Research & Management Science*. DOI: 10.1007/978-1-4419-0176-7_3.

Zhou, Y., & Talburt, J. R. (2011). Entity identity information management (EIIM). *MIT International Conference on Information Quality*, 237-341.

Bibliographic information of this paper for citing:

Ye, Yumeng, & Talburt, John R. (2018). The Effect of Transitive Closure on the Calibration of Logistic Regression for Entity Resolution. *Journal of Information Technology Management*, 10(4), 1-11.