Multiple Instance Learning for Group Record Linkage

Zhichun Fu[1], Jun Zhou[1], Peter Christen[1], and Mac Boot[2]

[1] Research School of Computer Science
College of Engineering and Computer Science
The Australian National University
Canberra ACT 0200, Australia
{sally.fu,jun.zhou,peter.christen}@anu.edu.au
[2] Australian Demographic and Social Research Institute
College of Arts and Social Sciences
The Australian National University
Canberra ACT 0200, Australia
mac.boot@anu.edu.au

Abstract. Record linkage is the process of identifying records that refer to the same entities from different data sources. While most research efforts are concerned with linking individual records, new approaches have recently been proposed to link groups of records across databases. Group record linkage aims to determine if two groups of records in two databases refer to the same entity or not. One application where group record linkage is of high importance is the linking of census data that contain household information across time. In this paper we propose a novel method to group record linkage based on multiple instance learning. Our method treats group links as bags and individual record links as instances. We extend multiple instance learning from bag to instance classification to reconstruct bags from candidate instances. The classified bag and instance samples lead to a significant reduction in multiple group links, thereby improving the overall quality of linked data. We evaluate our method with both synthetic data and real historical census data.

Keywords: Multiple instance learning, record linkage, entity resolution, instance classification, historical census data.

1 Introduction

Within many organisations, data are collected from various sources and through different channels, and they are stored in databases with different structures and formats. As organisations collaborate, data often need to be exchanged and integrated. The objective of such data integration is to identify and match all records that correspond to the same real-world entity, such as the same customer, patient, or taxpayer [10]. Record linkage (also known as data matching or entity resolution) is a key step to effectively mine rich information that is not available in a single database. This technology has been used in many areas, such as

P.-N. Tan et al. (Eds.): PAKDD 2012, Part I, LNAI 7301, pp. 171–182, 2012.

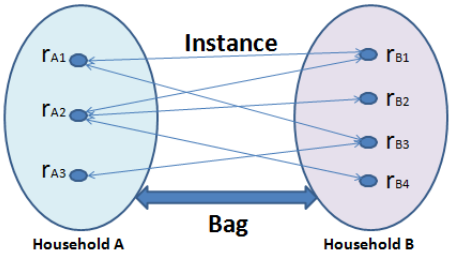

Fig. 1. An example of group (household) record linkage, and the corresponding MIL setting. Links between individual records correspond to instances while a bag is made of all links between the records in two groups.

electronic health record systems, the retail industry, business analytics, fraud detection, demographic tracking, and government administration [10].

As one application of record linkage, linking of historical census records across time can greatly enhanced their values by, for example, enabling tracking of households and providing new insights into the dynamic character of social, economic and demographic changes. In recent years, researchers have tried to link records between census datasets using automatic or semi-automatic methods [3,15,19,21]. Unfortunately, these attempts have not been not very successful in linking records that correspond to individuals in a household [20].

Several reasons make the linking of historical census records a challenging undertaking. First, the quality of historical census data is poor, because large amounts of errors and inaccurate information have been introduced during the census collection and digitisation processes [19]. Second, a large portion of records contain the same or similar values. It is not uncommon to find different people with the same name, the same age, and living in the same street in one dataset. Third, the structure of households and their members can change significantly between two censuses (which were normally collected every five or ten years). Therefore, simply comparing individual records does not lead to reliable linkage outcomes. Considering household information in the linkage process can help overcome this challenge.

In this research, we tackle the problem of linking individual records and households in historical census data. A household link will likely contain several links between individual record pairs for its household members. If two households are matching, at least one of their record links has to be a match. On the contrary, if two households are not matching, none of their record links shall be matched. This is a typical multiple instance learning (MIL) setting. MIL is a supervised learning method proposed by Dietterich et al. [9]. In MIL, data are represented as bags, each of which contains some instances. In a binary classification setting, a positive bag contains both positive and negative instances, while a negative bag only consists of negative instances. In the training stage, the class labels are only available at the bag-level but not at the instance-level. The goal of MIL is to learn a classifier which can predict the label of an unseen bag. When applying

MIL to the group record linkage problem, group links are treated as bags, and record links become the instances in these bags. A model can then be learned to classify a group link as a match or non-match. Figure 1 shows an example of group linking and its relationship to the MIL setting.

Because an individual record in one census dataset has generally a high similarity with several records in different households in another dataset, a household in one census dataset is often linked to different households in another dataset. Although such results can be helpful, e.g. in generating family trees, social scientists are often interested in tracking the majority of household members as a whole entity over time [20]. This suggests one-to-one household matches are needed. To reduce the number of multiple household matches, we can employ a group linking method [17,18], which generates a household match score for each household pair. Then the household pairs with the highest match score are selected as the final match results. Such an approach requires the detection of all matched record pairs in a household, which is equivalent to classifying instances within a bag as matches or non-matches. This is a problem that has not been adequately addressed in MIL research [16]. In traditional MIL methods [4,14], when instance selection is concerned, only the optimal positive instances are explored, whilst no explicit instance classification solution has been given. Therefore, there is a gap between MIL and its application to group record linkage.

We extend the above mentioned MIL methods to instance level classification by grouping negative instances from the training set with an instance to be classified. This transforms the instance into a bag. We can then employ the bag-level classification model for explicit instance classification. We show that this method can effectively classify both household and record links.

This paper makes two contributions. First, we extend the MIL method to instance classification via bag reconstruction. Second, we propose a practical solution to linking households between historical census datasets by group linkage using MIL. Our method is general in nature and it can be applied to other record linkage applications that require groups of records rather than individual records to be linked.

2 Related Work

In recent years, many methods have been developed for record linkage in the fields of machine learning, data mining and database systems [10]. Among them, supervised learning has been intensively investigated. It uses labelled record pairs with known match status (match or non-match) to learn a classification model. Bilenko et al. [2] proposed a solution based on support vector machines (SVM) [23] to compute the similarity between strings. Alternatively, Christen [6] has constructed inputs for a SVM using a pre-selection step which retrieves record pairs that with high confidence correspond to matches or non-matches. These pairs then become the positive and negative training samples for a SVM classifier. This method can be considered as a combination of supervised and un-supervised techniques.

Group record linkage methods have been developed to process groups rather than individual records [18]. On et al. [17] defined group similarity from two aspects, the similarity between matched record pairs and the fraction of matched record pairs between two groups of records. A group similarity can then be calculated using a maximum weight bipartite matching.

Multiple instance learning is a paradigm of machine learning that deals with a collection of data called *bags*. The original work by Dietterich et al. [9] attempted to recover an optimal axis-parallel hyper-rectangle in the instance feature space to separate instances in positive bags from those in negative bags. Departing from this model, several researchers have extended the framework, such as MI-SVM [1], DD-SVM [5], SMILE [24], MILES [4] and MILIS [14].

Among these works, we are particularly interested in the Multiple Instance Learning with Instance Selection (MILIS) method because it allows efficient and effective instance prototype selection for target concept representation [14]. This is an important property for (historical) census record linkage, which works on potentially large numbers of households and their records, and contains significant amounts of uncertainty because of low data quality.

MILIS is an extension of MIL using an embedded instance selection (MILES) method [4]. The general idea of these two methods is to map each bag into a feature space defined by selected instances, which is based on bag-to-instance similarity. It generates a feature vector for each bag, whose dimension is the number of selected instances. In this manner, the MIL problem is converted into a supervised learning problem, for which a SVM can be used for classification.

The major difference between MILES and MILIS methods is on the instance selection step. In MILES, all instances in the training set are used for feature mapping, then important features are selected by a 1-norm SVM. Because the total number of instances in a training set may be very large, MILES can be very time consuming. MILIS, however, only selects one instance prototype (IP) from each bag for the embedding. It generates a feature space with much smaller dimension than MILES. The selection of IPs is done through a two-step optimisation framework, which updates IPs and a SVM classifier iteratively.

3 Group Linkage Using Multiple Instance Learning

In this section, we introduce a group record linkage method based on multiple instance learning. Here, we treat a group link as a bag and its record links as instances in a bag, as shown in Figure 1. As mentioned before, group linking requires prediction on whether or not two records match, which is equivalent to instance classification. Therefore, we extend the MILIS algorithm so that a single instance can be grouped with negative instances in the training set to create a new bag. Then the bag can be classified using the learned bag-level classifier.

3.1 Instance Selection and Classifier Learning

To commence, we give formal definitions of the notion used in the method. Let $\mathcal{B}^+ = \{B_1^+, \ldots, B_{n^+}^+\}$ be a set of positive bags, $\mathcal{B}^- = \{B_1^-, \ldots, B_{n^-}^-\}$ be a set

of negative bags, and $n = n^+ + n^-$ be the total number of bags in the training set. A bag B_i contains m_i instances denoted by $\mathbf{x}_{i,j}$ for $j = 1, \ldots, m_i$, with the value for m_i varying from bag to bag. Each instance $\mathbf{x}_{i,j}$ is associated with a label $y_{i,j} \in \{1, -1\}$ that is not directly observable in the MIL setting, with $y_{i,j} = 1$ corresponding to a match and $y_{i,j} = -1$ to a non-match. The purpose is, therefore, to predict the binary label value $y_i \in \{1, -1\}$ for a novel test bag $B_i = \{\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,k}\}$, and $y_{i,j}$ for an instance $\mathbf{x}_{i,j}$.

Following the idea of instance-based embedding in [4] and instance prototype selection in [14], we generate bag-level feature representation using the similarity between a bag and an instance

$$s(B_i, \mathbf{x}) = \max_{\mathbf{x}_{i,j} \in B_i} \exp\left(-\gamma ||\mathbf{x}_{i,j} - \mathbf{x}||^2\right), \tag{1}$$

where γ is a feature mapping parameter that controls the similarity. Then a bag can be represented as an n-dimensional vector

$$z_i = [s(B_i, \mathbf{x}_1^*), \ldots, s(B_i, \mathbf{x}_i^*), \ldots, s(B_i, \mathbf{x}_n^*)], \tag{2}$$

where $\mathbf{x}_i^*$ are the prototype instances selected from the training set.

As proposed in [14], instance prototypes can be generated by selecting the least negative instance from each positive bag and the most negative instance from the negative bag. This requires modelling of the distribution of negative instances, and computing the probability that an instance has been generated from the negative population. Given an instance $\mathbf{x}$ and its k-nearest negative instances from the negative bags X_k^-, the likelihood of $\mathbf{x}$ being negative is

$$p(\mathbf{x}|X^-) = \frac{1}{Z} \sum_{j=1}^{k} \exp\left(-\beta ||\mathbf{x} - \mathbf{x}_j^-||\right), \tag{3}$$

where $\mathbf{x}_j^- \in X^-$ is the j^{th} nearest negative neighbour of $\mathbf{x}$, Z is a normalisation factor, and β is a parameter to control the contribution from training samples. We then select the instance with the lowest likelihood value from each positive bag as the positive instance prototypes (PIPs), and the instance with the highest likelihood value from each negative bag as negative instance prototypes (NIPs). These PIPs and NIPs form the set of instance prototypes (IPs) used in the feature mapping. Using Equations 2 and 3, we can represent bags in the training set in vector form, and then train a SVM classifier by solving the following unconstrained optimisation problem:

$$\min_{\mathbf{w}} \frac{||\mathbf{w}||^2}{2} + C \sum_i \max(1 - y_i(\mathbf{w}^T \mathbf{z}_i), 0), \tag{4}$$

where $y_i \in \{1, -1\}$ is the label for bag i, $\mathbf{w}$ is a set of parameters that define a separating hyper-plane, and C is the regularisation parameter [23].

3.2 Instance Classification

Both MILES and MILIS can find the most positive instance in a positive bag. This is achieved by selecting an instance in the bag that has the lowest likelihood value using Equation 3, because a positive bag should contain at least one positive instance. However, when it comes to the situation where a bag contains more than one positive instance, neither method provides an explicit solution to finding all the positive instances. Although a threshold may be set for decision, with instances whose likelihood is higher than the threshold classified as positive, and visa versa, it is practically difficult to find an appropriate threshold.

Here we propose a method for instance classification by bag reconstruction. We treat each instance in a positive bag as a seed, and group the instance with negative instances to create new bags. Then we apply the trained bag-level classifier to these new bags. If a new bag is classified as positive, then the seed instance is classified as positive. Otherwise, it is classified as negative. This method is based on the fact that if a seed is negative, the reconstructed bag consists of negative instances only, and thus will be classified as negative. Otherwise, the new bag contains one positive instance, therefore, is very likely to be classified as positive.

We have adopted two strategies for the bag reconstruction, **Random** and **Greedy**, to cope with multiple positive instances in a candidate bag. The first strategy randomly selects negative instances from the training set and groups them with the seed. Therefore, both the random negative instances from the training set and the seed instance contribute to the embedding step in MIL. The second strategy is built on top of the random option. With randomly selected negative instances, a greedy algorithm is adopted which reconstructs new bags and predicts the label of the newly added instance simultaneously. This guarantees not only the seed, but also the negative instances in the candidate bag, contribute to the embedding step. For each instance $\mathbf{x}$ in the candidate bag, we compute its Hausdorff distance to a bag G that contains NIPs $\mathbf{x}_i^{*-}$ only:

$$d(G, x) = \min_{\mathbf{x}_i^{*-} \in G} ||\mathbf{x} - \mathbf{x}_i^{*-}||^2 \tag{5}$$

Using this distance measure, we can get the similarity between an instance and the negative instances in G. By ranking the distances, we can construct a new bag by sequentially adding into the bag an instance with the lowest distance among the rest of the instances in the candidate bag. Evaluating the new bag using the bag-level SVM classifier, we can get the label of the newly added instance. For a candidate bag that contains both positive and negative instances, initially, the added instances are negative. Therefore, the bag is predicted as negative. When the prediction becomes positive after a new instance is added, the new instance is classified as positive. We then replace the positive instance with an instance that has a larger distance, and re-evaluate the new bag. This process continues until all instances in the candidate bag have been traversed. We summarise this strategy in Algorithm 1.

Algorithm 1. Instance Classification using Greedy Bag Reconstruction

Input:
- A set $\mathcal{B}^-$ containing all negative bags in the training set
- A bag G containing all NIPs
- A candidate bag B_i that contains m_i instances $\mathbf{x}_{i,j}$ for $j = 1, \ldots, m_i$
- Trained bag-level SVM model Φ
- An empty bag $\tilde{B}$

Output:
- Labels $y_{i,j} \in \{1, -1\}$ for instances $\mathbf{x}_{i,j} \in B_i$, for $j = 1, \ldots, m_i$

1: Randomly sample negative instances from $\mathcal{B}^-$, and add them into $\tilde{B}$
2: **For** $\mathbf{x}_{i,j} \in B_i$ **do**
3: Compute Hausdorff distance $d(G, \mathbf{x}_{i,j})$ using Equation 5
4: Sort $d(G, \mathbf{x}_{i,j})$ for $j = 1, \ldots, m_i$
5: Find $\mathbf{x}_{i,j}$ with the minimum $d(G, \mathbf{x}_{i,j})$ in B_i
6: Add $\mathbf{x}_{i,j}$ into $\tilde{B}$. Remove $\mathbf{x}_{i,j}$ from B_i
7: Classify $\tilde{B}$ using Φ
8: **If** $\tilde{B}$ is negative
9: $y_{i,j} = -1$
10: **Else**
11: $y_{i,j} = 1$. Remove $\mathbf{x}_{i,j}$ from $\tilde{B}$
12: **Goto** step 5

3.3 Group Record Linkage

The MIL step may generate a number of false positive bags. In the context of group record linkage, this means that a group in one dataset is possibly matched to several groups in another dataset. For applications such as linking households in (historical) census data, a one-to-one linkage of groups is often required, e.g., to track the majority of members of a household across time. We therefore use the group linkage method proposed by On et al. [18] to reduce the number of multiple matches between groups. This method computes a similarity score between two groups, which is based on the number of record pairs that have been matched between two groups and the total number of records in the two groups. This is equivalent to selecting a bag that has been classified as positive in the MIL step, and using the instance labels to compute a similarity score for this bag. In [18], the similarity is calculated using the following normalised weight of the matched individual record pairs in the two groups:

$$\mathbb{S}_{i,j} = \frac{\sum_{(r_a, r_b) \in M} sim(r_a, r_b)}{m_i + m_j - |M|}, \tag{6}$$

where M is the set of record pairs matched between groups H_i and H_j, r_a and r_b are the records in the two groups, and m_i and m_j are the number of records in the two groups. The set of all links between H_i and H_j is the bag, and the link between r_a and r_b is one instance in this bag. Therefore, the similarity function $sim(r_a, r_b)$ can take on the label predicted by the MIL model, i.e. $sim(r_i^a, r_j^b) = 1$ for matched record pairs and $sim(r_i^a, r_j^b) = -1$ for non-matched pairs. This approach reduces the group linking problem to computing the Jaccard index between two groups [22]. A final set of matched groups is then extracted by selecting the group links with the highest similarity value $\mathbb{S}_{i,j}$ among all pairs of groups. When several group links generate the same highest value, all of them

are considered as matches. Thus, the final output may still contain multiple links per group, but a much smaller number of them.

4 Experiments and Evaluation

We performed experiments on one synthetic dataset and six real census datasets using both the MILES and MILIS methods for the multiple instance learning step. For the implementation of MILES, we have used the MOSEK[1] system to solve the linear programming formulation in the one-norm SVMs. To train the MILIS algorithm, we have used LIBLINEAR [11]. The SVM regularisation parameter C was set using grid search on the training data. For Equation 3, we set $K = 10$ which is the same as in [14]. The feature mapping parameter γ in Equation 1 and the scale parameter β for the likelihood estimation in Equation 3 are both set to 1. For bag reconstruction in instance classification for the census data experiments, we have grouped a seed with 5 random negative instances. This is based on the fact that by average, a bag in the census datasets contains 5.65 instances, as can be calculated from Table 2.

For comparison purpose, we have implemented an alternative solution for bag and instance classification based on the group linkage method proposed by On et al. [18]. This method computes the sum of the similarity scores for each record pair, and then separates pairs into matches and non-matches by comparing the similarity sum with a threshold parameter ρ. The decision on the optimal ρ can be made based on the trade-off between the number of household pairs with multiple matches or unique matches. The matched households are then generated by grouping all matched record pairs that belong to the same matched household.

4.1 Synthetic Data Results

We have conducted experiments on synthetic data to evaluate the effectiveness of our instance classification method. The synthetic data generation follows the method in [14]. We randomly generated 1,000 positive instances and 5,000 negative instances, with each class generated from two Gaussian distributions. Then we constructed 50 positive bags by random sampling from both positive and negative instances, and 50 negative bags sampling from negative instances only. The number of instances in each bag is also randomly selected between 1 and 10. In this way, both bag and instance labels are known.

We split these bags into a training and a testing set, each containing 25 positive and 25 negative bags. We then trained bag-level classifiers using both the MILES and MILIS methods, and used them to classify instances in the testing set. This test is repeated 500 times over random partitions. The results show that the bag reconstruction method for instance classification presented in Section 3.2 is very effective. The random bag reconstruction method has achieved

[1] http://www.mosek.com

Table 1. Number of records and households in the historical census datasets

	1851	1861	1871	1881	1891	1901
Number of records	17,033	22,429	26,229	29,051	30,087	31,059
Number of households	3,295	4,570	5,575	6,025	6,379	6,848

Table 2. Number of bags and instances extracted from the historical census datasets

	1851–1861	1861–1871	1871–1881	1881–1891	1891–1901
Number of instances	2,104,171	2,200,876	2,459,272	3,043,786	3,318,738
Number of bags	325,921	441,355	472,239	494,270	588,436

an accuracy of $92.03 \pm 2.21\%$ using the MILES model, and $92.32 \pm 2.43\%$ using the MILIS model, while the greedy extension has achieved $92.89 \pm 2.89\%$ and $95.50 \pm 2.47\%$ on MILES and MILIS, respectively.

4.2 Historical Census Data Results

We used six census datasets from the district of Rawtenstall in the United Kingdom that were collected in ten-year intervals from 1851 to 1901. These census data contain twelve attributes per record, including the address, first and family name, age, gender, relationship to head, industry (occupation), and place of birth of each individual[2]. Because these data are of low quality, we have cleaned and standardised them using the *Febrl* data cleaning and record linkage system [7]. Details of this step can be found in Fu et al. [12]. Table 1 shows the number of records and households in each dataset.

The record level linkage was also conducted using *Febrl*. Instead of comparing all possible record pairs between two datasets, we used a traditional blocking technique combined with a Double-Metaphone encoding technique to index (block) the datasets [8]. We used a variety of approximate string comparison functions to calculate the similarity between individual record pairs following the approach given by Fu et al. [13]. The similarity scores calculated for a record pair were concatenated into a vector and then used in the MIL classification step.

We have manually labelled 1,000 household links from the 1871 and 1881 datasets, consisting of 500 matched and 500 non-matched households. To show the performance of the MILES and MILIS methods on household link classification, we performed 100-fold cross validation on the randomly split labelled data, with half used for training and half for testing.

Both the MILES and MILIS methods show similar performance, achieving $84.54 \pm 1.33\%$ and $83.75 \pm 1.34\%$ accuracy on household link classification, respectively. When efficiency is concerned, MILIS shows superior performance than MILES. The MILES method took 29.22 ± 6.37 seconds for training, and

[2] www.uk1851census.com

Table 3. Number of positive bags and instances classified in the different pairs of historical census datasets using the different methods described in this paper

	1851–1861	1861–1871	1871–1881	1881–1891	1891–1901
MILES-bag	7,728	9,644	9,705	9,650	12,583
MILIS-bag	8,832	11,369	9,870	9,175	11,282
Group-linkage-bag	47,249	50,494	49,306	48,212	50,058
MILES-random-instance	22,439	22,478	23,329	27,577	29,065
MILIS-random-instance	20,431	20,236	20,680	23,914	24,410
MILES-greedy-instance	22,063	21,771	23,170	27,019	28,987
MILIS-greedy-instance	20,738	21,436	22,228	25,050	24,872
Group-linkage-instance	67,122	67,340	65,528	65,595	67,483
After result fusion	775	1099	1484	1620	1689

0.88 ± 0.03 seconds for testing, while MILIS only took 2.17 ± 0.10 and 0.25 ± 0.04 seconds for each task. We did not evaluate the instance classification performance because the true record pair labels were not available to us.

In the next experiment, we re-trained the MILES and MILIS models using all the labelled data, and then classified all household and record links from any pair of consecutive census datasets, e.g. 1851 with 1861, 1861 with 1871, and so on. Because we were mainly interested in finding record matches in matched households, the instance classification was only performed on positively classified bags. As shown in Table 3, MILES and MILIS showed mixed performance on the bag-level classification, each having generated more positive bags than the counterpart on some datasets. By comparing the number of matched households with the total number of households in each census dataset (see Table 1), one can observe that the results contain multiple matches. This is expected because of two reasons. First, a household may split into several households, for example, due to the move-out of grown-up children, or two households might merge when widowed individuals form a new household. Second, there are many similar record pairs among different households, which may have generated false positive results. On the instance-level classification, the MILES-based models have consistently generated more positive instances than the MILIS-based models. The random bag reconstruction method, on the other hand, has achieved performance close to that of the greedy bag reconstruction method.

From Table 3, it can be observed that the group linkage method developed by On et al. [18] has generated many more household and record matches, i.e. more positive bags and instances, than the proposed MIL based methods. Statistics show that the MILES and MILIS based methods can reduce the number of matched bags in average between 79.98% and 79.40% respectively, when compared against the group linkage method described by On et al. [18]. Please note due to the lack of ground truth on household and record pair links, we have not used traditional measures such as accuracy and F-score for evaluation purposes.

We next applied the group linkage method introduced in Section 3.3 to reduce the number of multiple household matches, i.e. where one household is matched with multiple households. Figure 2 shows the performance of the proposed meth-

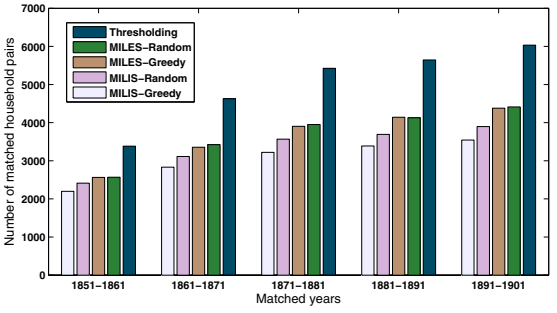

Fig. 2. Household matching results after group linkage step

ods and the thresholding method in [18]. The results indicate that the thresholding method generates the highest number of matches, followed by the MILES-based methods. The MILIS and greedy bag reconstruction combination has generated the smallest number of matches for all dataset pairs, which makes it the most reliable option in finding household matches between census datasets.

Finally, we performed results fusion so as to let the proposed methods vote for the most consistent household matches. This was performed by selecting household matches where all four options, i.e. MILES-random, MILES-greedy, MILIS-random, and MILIS-greedy, have agreed upon in their decision. These are the most reliable household matches that can be presented to researchers for further analysis. The last line in Table 3 shows the number of household matches after this fusion process.

5 Conclusion

We have introduced a group record linkage method based on multiple instance learning (MIL), and evaluated this method on real historical census data. In this method, group links are considered as bags and associated record links are treated as instances, with only the bag-level labels provided. The multiple instance learning paradigm has provided the group linkage problem with a suitable supervised learning tool to classify groups, even if the labels of record links are not available. We have shown the effectiveness of the proposed method on both synthetic and real historical census data from the UK.

In the future, we plan to extend the instance classification work so that instances selected for bag reconstruction better characterise the data distribution, and we will investigate approaches that allow linking records and households across several census datasets in an iterative fashion. We will also apply our method to other applications with a similar setting, such as bibliographic databases.

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: NIPS, Vancouver, Canada, pp. 561–568 (2003)

2. Bilenko, M., Mooney, R.J.: Adaptive duplicate detection using learnable string similarity measures. In: ACM KDD, Washington, DC, pp. 39–48 (2003)
3. Bloothooft, G.: Multi-source family reconstruction. History and Computing 7(2), 90–103 (1995)
4. Chen, Y., Bi, J., Wang, J.: MILES: Multiple-instance learning via embedded instance selection. IEEE TPAMI 28(12), 1931–1947 (2006)
5. Chen, Y., Wang, J.Z.: Image categorization by learning and reasoning with regions. Journal of Machine Learning Research 5 (2004)
6. Christen, P.: Automatic Training Example Selection for Scalable Unsupervised Record Linkage. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 511–518. Springer, Heidelberg (2008)
7. Christen, P.: Development and user experiences of an open source data cleaning, deduplication and record linkage system. ACM SIGKDD Explorations 11(1), 39–48 (2009)
8. Christen, P.: A survey of indexing techniques for scalable record linkage and deduplication. IEEE TKDE (2011)
9. Dietterich, T.G., Lathrop, R.H., Lozano-Perez, T.: Solving the multiple-instance problem with axis-parallel rectangles. Artificial Intelligence 89, 31–71 (1997)
10. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering 19(1), 1–16 (2007)
11. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: LIBLINEAR: A library for large linear classification. JMLR 9, 1871–1874 (2008)
12. Fu, Z., Christen, P., Boot, M.: Automatic cleaning and linking of historical census data using household information. In: IEEE ICDM Workshop on DDDM (2011)
13. Fu, Z., Christen, P., Boot, M.: A supervised learning and group linking method for historical census household linkage. In: AusDM (2011)
14. Fu, Z., Robles-Kelly, A., Zhou, J.: MILIS: Multiple instance learning with instance selection. IEEE TPAMI 33(5), 958–977 (2011)
15. Fure, E.: Interactive record linkage: The cumulative construction of life courses. Demographic Research 3, 11 (2000)
16. Li, F., Sminchisescu, C.: Convex multiple instance learning by estimating likelihood ratio. In: NIPS (2010)
17. On, B.-W., Elmaciogl, E., Lee, D., Kang, J., Pei, J.: Improving grouped-entity resolution using quasi-cliques. In: IEEE ICDM, Hong Kong, pp. 1008–1015 (2006)
18. On, B.-W., Koudas, N., Lee, D., Srivastava, D.: Group linkage. In: IEEE ICDE, Istanbul, Turkey, pp. 496–505 (2007)
19. Quass, D., Starkey, P.: Record linkage for genealogical databases. In: ACM KDD Workshop, Washington, DC, pp. 40–42 (2003)
20. Reid, A., Davies, R., Garrett, E.: Nineteenth century Scottish demography from linked censuses and civil registers: a 'sets of related individuals' approach. History and Computing 14(1+2), 61–86 (2006)
21. Ruggles, S.: Linking historical censuses: a new approach. History and Computing 14(1+2), 213–224 (2006)
22. Tan, P., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Addison-Wesley (2005)
23. Vapnik, V.: The Nature of Statistical Learning Theory. Springer (1995)
24. Xiao, Y., Liu, B., Cao, L., Yin, J., Wu, X.: SMILE: A similarity-based approach for multiple instance learning. In: IEEE ICDM, Sydney, pp. 309–313 (2010)