Heterogeneous Committee-Based Active Learning for Entity Resolution (HeALER)

Xiao Chen(✉), Yinlong Xu, David Broneske, Gabriel Campero Durand, Roman Zoun, and Gunter Saake

Otto-von-Guericke-University of Magdeburg, Magdeburg, Germany
{xiao.chen,yinlong.xu,david.broneske,campero,roman.zoun,saake}@ovgu.de

Abstract. Entity resolution identifies records that refer to the same real-world entity. For its classification step, supervised learning can be adopted, but this faces limitations in the availability of labeled training data. Under this situation, active learning has been proposed to gather labels while reducing the human labeling effort, by selecting the most informative data as candidates for labeling. Committee-based active learning is one of the most commonly used approaches, which chooses data with the most disagreement of voting results of the committee, considering this as the most informative data. However, the current state-of-the-art committee-based active learning approaches for entity resolution have two main drawbacks: First, the selected initial training data is usually not balanced and informative enough. Second, the committee is formed with homogeneous classifiers by comprising their accuracy to achieve diversity of the committee, i.e., the classifiers are not trained with all available training data or the best parameter setting. In this paper, we propose our committee-based active learning approach HeALER, which overcomes both drawbacks by using more effective initial training data selection approaches and a more effective heterogenous committee. We implemented HeALER and compared it with passive learning and other state-of-the-art approaches. The experiment results prove that our approach outperforms other state-of-the-art committee-based active learning approaches.

Keywords: Entity resolution · Query-by-committee-based active learning · Learning-based entity resolution · Record linkage

1 Introduction

Entity resolution (ER) is the task of identifying digital records that refer to the same real-world entity [6]. The classification step in an ER process can be considered as a binary classification problem [9]. Supervised learning can be

This work was partially funded by the DFG [grant no.: SA 465/50-1], China Scholarship Council [No. 201408080093] and Graduiertenförderung des Landes Sachsen-Anhalt.

T. Welzer et al. (Eds.): ADBIS 2019, LNCS 11695, pp. 69–85, 2019.
https://doi.org/10.1007/978-3-030-28730-6_5

adopted to solve this problem. However, in order to reach a satisfactory accuracy, a high amount of training data has to be provided, which is usually not available, and has to be labeled by domain experts. The training dataset labeled by domain experts for ER tasks is even more difficult to get than a normal classification problem, since for each labeling, experts have to work on both records of pairs and all their attribute pairs to make the final decision. Therefore, reducing the number of required training data is of great importance for ER.

As a means to reduce human effort, active learning (AL), which is a specific branch of machine learning (ML), is proposed to deal with this problem. Compared to a normal one-off ML process, an AL process is interactive and iterative. It reduces the number of required training data to achieve a desired accuracy by querying experts to label only the most informative data for each iteration and adding these into the training data (those pairs that are intrinsically difficult to classify based on available features are informative data). Then classifiers are retrained on updated training data and after each iteration, the stopping criteria are checked to see whether more iterations are required. So far, there have been different AL approaches proposed, which differ on the strategies to choose the most informative data. *Query by committee (QBC)* is an effective approach that has been successfully applied to many applications [16]. It gets the most informative data by selecting those data that result in the most disagreement in the votes of a committee of multiple classifiers [25]. However, applying QBC approaches for AL-based entity resolution (AL-ER) problems faces two challenges:

Diversified Committee: The key challenge to make QBC work in common is to generate a diversified committee, which can insightful voting disagreements so that the informativeness of data can be represented and distinguished [16] [14]. In order to achieve the diversity, for the vast majority of proposed AL approaches, they consider how to get multiple models with only one single type of classification algorithm. So far, several methods have been proposed for the diversity purpose, such as query by bagging, query by boosting [15]. However, for all those ensemble-based approaches, the accuracy of each model is compromised to get this diversity. For instance, in the bagging approach, the initial training dataset is divided into several smaller subsets, then different models are built based on different subsets [15]. Those trained models cannot be expected to achieve such accuracy as the model trained on the whole training dataset. Besides, nowadays, data is also quite variable in their types and there is no universally best model for all types of data. If a system completely relies on a single type of model, accuracy could not be acceptable for the worst cases.

Imbalanced ER Classification: The second challenge specialized for an AL-ER solution is the generation of the initial training dataset. The binary classification task for ER is a special task because of the imbalance of its two groups. In our real world, there are much fewer match pairs than non-match pairs, e.g., for the well-known Scholar-DBLP dataset, the imbalance ratio is 1 match but 3273 non-matches [28]. If the initial training dataset is randomly selected from all candidate pairs, the possibility to contain match pairs would be quite low, which may lead to a very low starting accuracy of trained models or even fail in training a model. Facing imbalanced data, oversampling and undersampling are commonly-used. However, except for their intrinsic shortcomings (overfitting for oversampling and discarding potentially useful data for undersampling [12]), they also contradict the goal of AL: saving labeling effort as much as possible. Facing both challenges, we propose in this paper a novel **He**terogeneous **A**ctive **L**earning **E**ntity **R**esolution (HeALER) solution. We specifically detail our contributions as follow:

- We design a specialized technique to generate the initial training dataset, which is suitable for the inherent class imbalance in ER;
- We propose to construct the AL committee with different types of classification algorithms, through which we can achieve diversity, accuracy and robustness requirements of a committee;
- We prototype our solution and evaluate it with two well-known ER benchmarking datasets, and comparing with passive ML and two state-of-the-art AL-ER approaches (ATLAS [27] and ALIAS [24]). The evaluation results show that HeALER is faster to converge and can reach a higher final F-measure, which also indicates that with fewer labels a satisfactory F-measure can be achieved.

The remainder of this paper is organized as follows: In Sect. 2, we introduce our HeALER approach. Subsequently, we evaluate our approach and discuss the experiment results in Sect. 3. Before we conclude and list future work in Sect. 5, we also compare our method to other related work in Sect. 4.

2 Heterogeneous Committee-Based Active Learning for Entity Resolution

In this section, we introduce our designed QBC AL method for ER, which is characterized by its initial training data selection approach and its heterogeneous committee. We start with a global picture of our approach in Sect. 2.1, then we represent our initial training data selection method, heterogeneous committee in the following sections.

2.1 The Global Workflow

Figure 1 represents the global workflow of our method HeALER. It is separated into two parts, the left green area describing the preparation steps, and to the side the light red area corresponds to the AL stage.

Preparation for Active Learning. As we can see from the left green area of Fig. 1, several preparation steps are required to start the ER process. At first, input data is preprocessed if necessary, which may include data cleaning, formatting, standardization. Afterwards, blocking is performed to omit unnecessary comparisons, which are obvious non-matches based on predefined blocking keys [5]. Then candidate pairs are generated based on the blocking result. Subsequently, for each attribute pair, one or more similarity functions are chosen

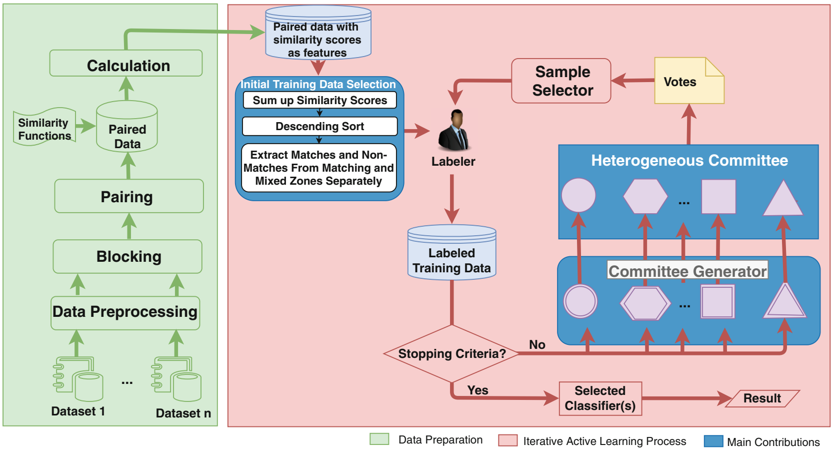


Fig. 1. The global workflow of HeALER

to best calculate similarities between each attribute pair in order to get similarity scores as features for the following learning-based classification step [4]. For the above-introduced steps, proper techniques should be employed based on ER task requirements, our contributions are reflected on the AL part, which will be briefly introduced next.

Iterative Active Learning Process. The first step of HeALER is to select pairs from the candidate pairs to be labeled by domain experts for an initial training dataset. As mentioned in Sect. 1, the classification step of ER is an imbalanced binary classification problem, i.e., there are much fewer match pairs than non-match pairs [9]. In order to reach a relatively high starting point with the initial training dataset, the training data is required to be balanced and informative. Balanced means the initial training dataset should contain sufficient percentages of match and non-match pairs, which is hard to achieve when one randomly picks pairs from the entire input data, since too high percentage of non-matches would be selected. Informative means the initial training data could involve useful information, which can benefit classifiers. The details how we achieve both goals will be introduced in Sect. 2.2.

Based on the initial training dataset, different classifiers are trained on them and then all classifiers together form the required committee. Notably, our classifiers are trained by different classification algorithms, which means our committee is heterogeneous. Compared to the majority of state-of-the-art QBC-AL approaches, our heterogeneous committee has the following advantages: First, the fundamental requirement - diversity of the committee - is achieved in a natural way without any other efforts. Second, each member of the committee is trained with the best or full ability without any compromise, which is more promising to provide a more accurate vote. Last, the committee analyzes training data and provides the result from multiple perspectives, no matter which kind of data the committee is facing, it can provide relatively stable and acceptable results. The methods to form our committee, including how to define the number of required classifiers and how to select classification algorithms as committee members, will be explained in Sect. 2.3.

After the committee is formed, they are employed to vote each pair from the unlabeled pool into match or non-match. The calculation of the disagreement of voting results for pairs will be firstly represented in Sect. 2.4. Then this process is iterated until the stopping criteria are reached.

2.2 Initial Training Dataset Generation

As explained in the last section, a good initial training dataset should be balanced and informative. In order to achieve both criteria, we analyzed a learning-based ER process. The resources that we have for the classification step are the candidate pairs and already calculated similarity scores for each attribute pair as features. Figure 2 is a histogram formed for the benchmarking bibliography dataset ACM-DBLP [13], which describes how the percentages of matching and non-matching pairs varies along with different similarity score levels. There are four attributes in this dataset, in total there are 16 similarity scores calculated as features (five similarity scores for the first three attributes: title, author, venue with different string similarity calculation functions; and one similarity score calculated for the attribute: publication year), each separate similarity score is normalized between zero and one, then the total similarity scores of all pairs should be between zero and sixteen by summing up all similarity scores. Based on this, we divided all candidate pairs into 15 groups and each group is an interval between *n* and *n+1* (*n* is from 0 to 15). As we can see from it, globally the whole pairs are located in three zones. For areas with the lowest similarity scores the vast majority of pairs are non-matching pairs (the non-matching zone). Then the percentage of matching pairs increases in relatively middle levels (the mixed zone), and for the last levels with highest similarity scores, the vast majority of pairs become matching pairs (the matching zone). Dealing with variable datasets, the concrete ranges of the three zones may vary, however, globally speaking, those three zones and their trends should be valid for almost all datasets.

From the perspective of balance, the difficulty for the imbalanced classification step of ER is to find a sufficient number of matching pairs, while non-match pairs are quite easy to get, because there are much more non-matching pairs than matching pairs in our real world. The percentages shown in the figure can indicate the difficulty to get matching and non-matching pairs. In order to get sufficient matching pairs, the matching zone has to be focused. In order to get sufficient non-matching pairs, both the non-matching zone and the mixed zone can be the candidates.

From the other perspective of being informative, those pairs that are intrinsically difficult to classify based on available features, can be considered as informative data, since the classifier would be significantly improved if informative pairs are labeled and added to help the classifier training. Hence, those error-prone pairs should be true match pairs with relatively *low* similarity scores and true non-match pairs with relatively *high* similarity scores. True match pairs with relatively *low* similarity scores should be located in the mixed zone, but it is not possible to get them, since the matching pairs account for very small percentages in the mixed zone. Therefore, for achieving both balance and informativeness, we have to pick matching pairs from the matching zone. On the other hand, for non-match pairs, true non-match pairs with relatively *high* similarity scores locate in the mixed zone, by combining the conclusion from above (the non-matching zone and the mixed zone for getting non-matching pairs from the perspective of balance), the mixed zone is the aiming zone for high quality non-matching pairs.

Based on the above considerations, we conclude our method to generate the initial training dataset for learning-based ER in the following way:

1. First, there can be many similarity scores calculated and for different attributes, values of similarity scores may vary much. Hence, it is difficult to look into each separate similarity score and judge the possibility based on them separately. Therefore, we calculate a total score of each pair by summing up all similarity scores of attributes.
2. Next, we sort all candidate pairs based on their total scores in descending order.
3. Last, we divide all sorted pairs into k groups, then we can get the initial training dataset by randomly picking $n/2$ number of pairs from the top k_1 groups (the matching zone) for getting sufficient matching pairs and $n/2$ number of pairs from the next k_2 groups (the mixed zone) for getting sufficient and informative non-matching pairs (n is the preset number of initial training data). There is no accurate method to determine which k, k_1, k_2 are the best. The following hypotheses can be used. If the ER problem is between two data sources and the linkage is one-to-one linkage, the highest number of matches is the number of records in the smaller dataset. This number can be used as the size of the matching and mixed zones. If the linkage is one to many, even many to many linkages, an information that can be used is the approximate percentage of matching pairs, then this can be the basis to locate the matching zone and the same percentage of pairs can counted for the mixed zone. If even the percentage of matching pairs is unknown, as a rule of thumb, 10 groups should be a good number to averagely divide all pairs with a proper blocking step, then the matching zone is the top group with highest similarity scores and the mixed zone corresponds to the second group for getting non-match pairs.

With the above-introduced strategy the interesting areas analyzed above are established. With the first top groups, we are able to get sufficient matching pairs, and with the next groups, sufficient and informative non-matching pairs can be obtained.

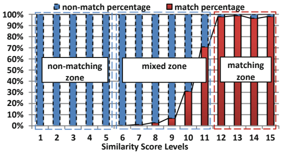


Fig. 2. Distribution of similarity scores

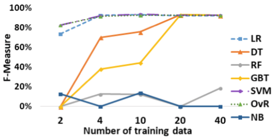


Fig. 3. F-measures of 7 classification algorithms

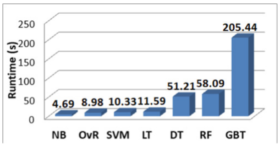


Fig. 4. Efficiency of 7 classification algorithms

2.3 Heterogeneous Committee

As introduced in Sect. 1, our committee is heterogeneous, which means that classifiers of the committee are trained with different classification algorithms. The method designation focuses on two aspects: how many classifiers and which classifiers to choose.

Generally speaking, our heterogeneous committee is allowed to contain any number of classifiers. Based on the result in [24], the performance of the classifier is not too sensitive to how many members a committee has and with four classifiers the aggregated accuracy is already satisfactory enough. On the one hand, each additional member in the committee means one more training process per iteration, which can heavily increase time needed for generating one round committee and has negative impact on efficiency. On the other hand, having more than four members for the committee achieves even lower accuracy [24]. Therefore, in our evaluation, four classifiers are generated to form the committee. Next, we present which candidate algorithms are suitable to be committee members. In general, we considered the following factors:

Accuracy with Little Training Data: The selected classifiers should have relatively high accuracy. Particularly, because the purpose of using AL is to reduce required human labeling efforts, we assume that for the training dataset, not so much training data is required to achieve high accuracy, which means that the selected classifiers should still work when only little training data is available. This is the main factor we use to choose classification algorithms.

Efficiency: Efficiency also requires consideration, since a learning-based classification is much more time-consuming than a simple threshold-based classification and such factor can be expected to have a large impact on the performance, as data grows.

Interpretability: Interpretability is also of great importance for choosing the learning algorithms, because we can use machine learning responsibly to ensure that our values are aligned and our knowledge is reflected [8].

We considered the following seven common binary classification algorithms: logistic regression (LR), decision tree (DT), random forest (RF), gradient-boosted tree (GBT), support vector machine (SVM), one-vs-rest logistic regression (OvR) and naive bayes (NB). In order to select classifiers for our committee, we evaluated their F-measures also on the same benchmarking bibliography dataset ACM-DBLP [13] used in Fig. 2 by using different sizes of initial training data. Figure 3 shows the results. SVM, OvR and LR have a satisfactory F-measure value even with only two training data pairs. NB and RF provide still a very low F-measure value even with 40 training data instances. NB classifiers are generative models that need prior probabilities. The probabilities are inaccurate for our case, because our initial training data is chosen by our proposed method in the last section, which normally generates relatively balanced training data. This state of training data does not conform to the test data [21]. Besides, it assumes that all features are independent [23]. However, our features are actually not independent, which may lead to the low F-measure for NB classifiers. The RF classifier cannot perform well, because it trains an ensemble of decision tree classifiers by splitting the training dataset into multiple subsets, then chooses subsets of features for each decision tree classifier [26]. This leads to a low F-measure especially when there is not enough training data. DT overall performs well except for the case with two training data pairs, in which DT classifier is not possible to be trained. GBT is in a similar situation as DT. However, its F-measure values are always lower than DT. We also evaluate the efficiency of all seven classification algorithms. All candidate pairs are divided into roughly two equal groups. Training data is generated by randomly picking four matching and non-matching pairs from the first group and then test data is the other entire group. The results are shown in Fig. 4. As we can see from it, results show that DT, GBT and RF need obviously more time than the other algorithms. NB runs the fastest, OvR, SVM and LR follows. However, all three tree-based classification algorithms DT, RF and GBT are quite slow. By combining the perspectives of interpretability and efficiency with the accuracy result, SVM, OvR, LR, and DT are selected to form our heterogeneous committee.

Above we provided guidelines on how to choose classification algorithms to form the heterogeneous committee. Facing different implementations of algorithms with different adopted libraries, the best choices of classification algorithms may change case by case.

2.4 Training Data Candidate Selection

After our heterogeneous committee is formed based on the above introduced approach, it is used to vote unlabeled pairs as matches or non-matches. Then those pairs with the most disagreement are those interesting pairs that we may select to be labeled by domain experts and added to the training dataset. The disagreement value of the voting results for pairs is calculated with the following equation:

$$Disagreement(pair) = \sum_{(a_m,a_n)\in committee)} Difference(result(a_m), result(a_n)) \quad (1)$$

where (a_m, a_n) are the combinations of results from any two classification algorithms from the committee and the *Difference(x, y)* function returns zero or one,

Table 1. Datasets used in experiments

Datasets	#Input	#Records in DBLP	#Records in ACM/Scholar	#Pairs (#match pairs)	#For training data selection	#For testing
ACM-DBLP2	2	2616	2294	21095 (2189)	10547	10548
Scholar-DBLP1	2	2616	64263	44999 (4351)	22500	22499

depending on whether x equals to y or not. With this equation, we sum up all the differences between any two combination of classification algorithms as the final disagreement value of the votes. However, the pair with a high disagreement value has also a high possibility that it is an outlier. If an outlier is selected and added to the training dataset, it will negatively impact the performance of classifiers. In order to reduce the possibility that outliers are selected, the random sampling proposed in [24] randomly picks the pair from the *top-n* pairs to alleviate the probability that an outlier is selected to be labeled, n can be set manually, such as 10, 20, 30.

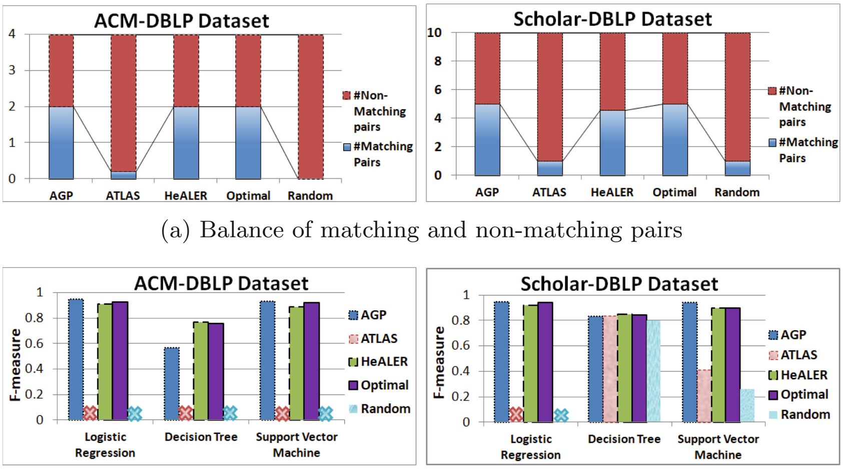
Then the training data is updated in the above-introduced way iteratively and after the iteration process is completed according to preset termination conditions, the committee or a specific classifier can be used to identify duplicates for any unlabeled data.

3 Evaluation

In this section, we evaluate HeALER from three aspects: first, we solely conducted experiments to evaluate the balance and accuracy of our initial training data selection method (Sect. 3.2). Second, we evaluate our heterogeneous committee and compare it to the passive learning, committees formed by ALIAS and ATLAS (Sect. 3.3). Last, we evaluate our entire HeALER approach against a ML process and two state-of-the-art QBC-AL approaches: ALIAS and ATLAS (Sect. 3.4). For all results, the accuracy is measured using F-measure.

3.1 Experimental Setting

Datasets: We evaluate HeALER on two commonly-used real-world datasets: ACM-DBLP and Scholar-DBLP citation datasets [13]. Both datasets include two parts, one part is from the DBLP citation database and the other one is from ACM or google scholar citation databases, respectively. All of them have four attributes, including title, authors, venue and publication year. In order to prepare data for HeALER, we have done the following steps based on the two original citation databases: We first preprocess both databases by removing stop words and null values. Then we generate blocking keys (the first five letters of the title) for each record. Subsequently, we join two database tables with the blocking key as the join attribute, so that we get all candidate pairs. Afterwards, similarity functions are performed on each attribute to get corresponding features.



(a) Balance of matching and non-matching pairs

(b) F-measure Comparison

Fig. 5. Initial training dataset selection approaches evaluation

For attributes "title", "author", we apply cosine, Jaccard, Jaro-Winkler, metric longest common subsequence, N-Gram, normalized Levenshtein and Sorensen-Dice similarity functions[1]. For attribute "venue", the Jaccard similarity function is used. For the last attribute "year", the similarity between two values is one or zero based on whether they equal or not. In this way, we obtained 16 features. For the preparation of our initial training dataset selection method, total similarity scores are calculated and appended to data as well. With the above introduced steps, for the ACM-DBLP dataset, we got 21095 pairs after blocking (including 2189 true match pairs). We randomly divide all pairs into two parts: the first half 10547 pairs as the first part form the dataset to select training data and the remaining pairs for testing. For the Scholar-DBLP dataset, we got 44999 pairs after blocking (including 4351 true match pairs). We also randomly separate it into two parts in the same way as the ACM-DBLP dataset. The details of datasets are summarized in Table 1.

Implementation Related: Since learning-based classification is much more time-consuming than threshold-based classification, we implemented HeALER with Apache Spark (version 2.4), which is a general framework supporting distributed computation, as a preparation for big data processing. However, this paper focuses only on the quality side of ER results. The classification algorithms used are implemented with Spark MLlib. The programming language is Scala with the version 2.11.12.

[1] Implemented by the Debatty library (version 1.1.0).

3.2 Initial Training Dataset Evaluation

Experimental Design. This experiment is to evaluate different strategies to select the initial training dataset by getting the average results over five runs. We use both datasets in Table 1. The tested initial dataset sizes are four and ten, which are proved to be the least to function selected classifiers (Fig. 3). The following strategies are evaluated:

Random Selection: It means we randomly select the required number of pairs.

Optimal Selection: The optimal selection means that training data is optimally balanced, i.e., because we have the ground truth for our datasets, we pick half matches and half non-matches from the unlabeled data. However, this is not practical, since before labeling, we have no idea which pairs are matches or non-matches. In [24], they selected initial training data in this unpractical way.

Initial Training Data Selection of ATLAS [27]: ATLAS ranks all pairs on their total similarity scores, then divides the whole pool to n groups (4 or 10 groups for two tested dataset sizes respectively), at last the initial training dataset is obtained by randomly selecting one data pair from each group.

Initial Training Data Selection of AGP [7]: In order to get both matching pairs and non-matching pairs, the initial training dataset of AGP is obtained by selecting half number of pairs with highest total similarity scores (2 or 5 pairs for two tested dataset sizes respectively) and the other half number of pairs with lowest total similarity scores (2 or 5 pairs for two tested dataset sizes respectively).

Initial Training Data Selection of HeALER: Our own method HeALER selects the initial training dataset in the way of the hypotheses described in Sect. 2.2. Since the linkage for ACM-DBLP dataset is one-to-one linkage, the highest number of matches is the number of records in the smaller dataset, i.e., 2294 records from ACM library. As the whole dataset is almost equally split to two datasets. Then the matches contained in the first dataset to select training data should be 1147. This number can be used to get the matching and mixed zones, i.e., two pairs randomly picked from the first 1147 pairs with the highest similarity scores, and two pairs randomly picked from the next 1147 pairs. Regarding the other dataset Scholar-ACM, it is not one-to-one linkage, but we know that the approximate percentage of its matching pairs is 10, therefore, we divide all pairs into 10 groups, and the first top group with the highest total similarity scores is the matching zone, where we randomly get 5 pairs, and the second group is the mixed zone, where we randomly get the rest 5 pairs.

We evaluate those above-introduced selection methods with balance and F-measure metrics. For the balance metric, how many matching and non-matching pairs in the training dataset is shown. For the F-measure metric, F-measures values are calculated by testing the classifiers trained on different training datasets with LG, DT, SVM classification algorithms respectively on the test dataset.

Results and Discussion. As can be seen from Fig. 5a, with random and ATLAS approaches, the training data selected is quite skewed, no sufficient matching pairs are picked, especially the random selection for the ACM-DBLP dataset selects no matching pairs, which may make the training data unusable, since some classifier algorithms cannot work with only one class of data for a binary classification problem. HeALER can achieve relatively balanced training data, but not as completely balanced as AGP and Optimal selection. The F-measures using LR, DT, and SVM calculated on the training data selected with different approaches are shown in Fig. 5b. Therein, the training data selected using ATLAS and the random approach works only for DT and SVM on the Scholar-DBLP dataset. For all other cases, no classifiers are successfully trained and used for the later test classification because of exceedingly skewed training data. The other three approaches work apparently better. With the training data they selected, it is always possible to complete the classification tasks using the trained classifiers. Particularly, HeALER outperforms AGP and the optimal case with DT due to the more informative training data, which makes the splitting closer to the truth. However, it achieves a bit lower F-measure for LR and SVM. By concluding the results, we can say that the quality of HeALER training data is high when the number of divided groups can be correctly defined. Otherwise, the AGP strategy can be applied to achieve acceptable F-measure.

3.3 Heterogeneous-Committee Evaluation

Experimental Design. This experiment is designed to specially evaluate our heterogeneous committee and compare it to other approaches (committees formed in [27] and [24] and passive learning to randomly pick pairs without basing on committees' decisions). Both datasets in Table 1 are used. We fix the initial training data selected by our own strategy for all approaches, which provides them fair and good starting points. And the strategy used to reduce the possibility to get outliers is fixed with the Sampling20 approach (It is evaluated as the best strategy by comparing Sampling10, Sampling20 and Sampling30 using the random sampling method (n is set as 10, 20 or 30) introduced in Subsect. 2.4. The evaluation results are omitted due to limited space). After each iteration of the AL process, the F-measure is calculated on the classification results of the test data in Table 1 obtained by using the DT classifier trained on the updated training datasets by each approach. The AL process terminates after 199 rounds. Each experiment is repeated three times to get the final average result. The details how different approaches perform are introduced as follows:

Passive Learning: This approach randomly picks pairs to be labeled by humans and added to the training dataset without relying on any committee votings.

ALIAS Committee [24]: ALIAS forms its committee by randomizing parameters while training classifiers with the selected algorithm. In our experiments for both datasets, SVM algorithm is used. We vary the value of its parameter for maximum number of iterations with 4, 6, 8 and 10. Then four classifiers are trained respectively and form its committee. In the ALIAS paper, for their experiments, they applied DT algorithm and varied the parameter where to split. However,

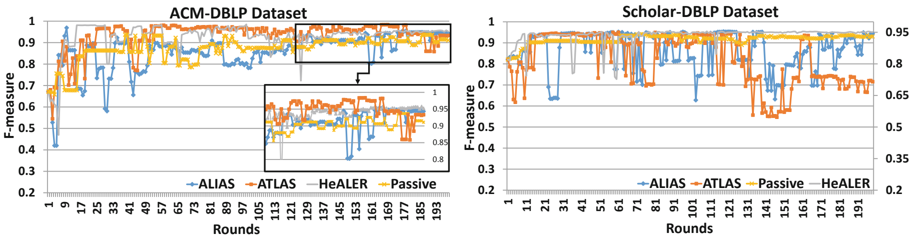


Fig. 6. Different committee comparison

as our implementation depends on the Spark MLlib, it is not possible to adjust this parameter. Therefore, we apply SVM algorithm for our experiments.

ATLAS Committee [27]: ATLAS partitions the training dataset to four subsets, then each subset of the training data is used to train its classifier to form its committee. The classification algorithm used here is the same as ALIAS: SVM for the purpose of comparison. For both datasets, each time 80 percent of pairs are randomly chosen to constitute the training dataset. Four subsets are required to get four classifiers of the committee.

HeALER Committee: As explained in Sect. 2.3, our heterogenous committee includes four classifiers, which are trained with SVM, OvR, LR, and DT algorithms, using the complete training dataset.

Results and Discussion. Figure 6 shows the comparison results of different committees and the passive learning. As we can see from the results of ACM-DBLP dataset, the F-measures of all approaches fluctuate much in the first 50 rounds, then becomes more stable later on. After about 140 rounds, our heterogeneous committee keeps F-measures higher than 0.92 and reaches its rough convergence. In contrast, ALIAS and ATLAS committees still cannot achieve their convergences till 199 runs. They show even less stable and lower results than passive learning. However, the highest F-measures they are able to reach during the experiments are much higher than passive learning, which proves the effectiveness of the committee to explore which are more informative pairs. Since the passive learning randomly chooses more pairs to be labeled, the informative pairs are hard to be selected to really cover the shortages of the classifiers. From the result of the other Scholar-DBLP dataset, we can get similar conclusions. Our heterogeneous committee converges already after about 110 rounds and keeps the F-measure 0.95 afterwards. ALIAS and ATLAS committees are far from their convergence even with 199 rounds. The passive learning works quite good for this dataset due to the high initial F-measure. However, it requires much labeling effort to improves its F-measure. To summarize the results, our heterogeneous committee shows its advantage in picking informative data to improve the F-measure of the classifier and reach the convergence with much less labeling efforts than passive learning, ALIAS and ATLAS committees.

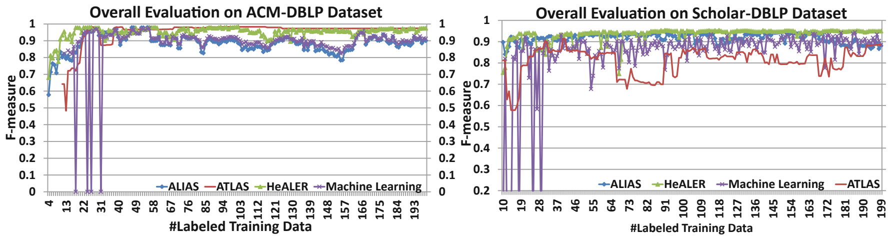


Fig. 7. Overall evaluation

3.4 Overall Evaluation and Comparison

Experimental Design. After we evaluate our initial training data selection approach and our heterogenous committee separately, in this section, we evaluate our entire HeALER approach by comparing the F-measures using a one-off ML approach, ALIAS and ATLAS approaches based on a same number of training data. For this overall evaluation, the same datasets are used as in the last two sections. The F-measures are all calculated on the classification results of the test data in Table 1 obtained by using the DT classifier. Except the normal machine learning approach, other approaches follow the iteration process of AL and terminate after 199 rounds. For the normal one-off ML approach, we randomly picked the corresponding number of training data of each iteration and calculate the F-measure of the test data using the DT classifier. ATLAS has no strategy to reduce the possibility to get outliers but it chooses the pair with the highest similarity value among all pairs with the highest disagreement value. Therefore, in the overall evaluation, for ATLAS, this approach choosing the pair with the highest similarity value is used. For HeALER and ALIAS, the sampling20 strategy is used as in the committee comparison experiment. The final result is averaged by three times' repetition (Fig. 7).

Results and Discussion. Figure 7 shows the comparison results of different AL approaches and a normal ML process. As we can see from the results of ACM-DBLP dataset, HeALER has the highest initial F-measure and keeps a F-measure around 0.9 with 20 or more training data. ALIAS and ML perform the worst and fluctuate their F-measures from the beginning to the end. ML starts to function stably with at least 33 labeled data and cannot significantly improve its F-measure when labeling more data. ATLAS starts to work with 10 labeled data and hardly varies its F-measure. The reason can be because its strategy always selects data with the highest total similarity score, weakens the effects of the disagreement values of data, and often chooses same data for different iterative rounds, which leads to changeless F-measure for several or even dozens of iterative rounds. Although it seems that ATLAS performs quite good, the results of the Scholar-DBLP dataset, in which ATLAS performs the worst, shows that ATLAS is not reliable, more research on the strategy of selecting the highest similarity score from the data with highest disagreement values is required. For the results of ALIAS, HeALER and ML on Scholar-DBLP, similar conclusions can be made. The results show that HeALER works better than the other compared approaches.

4 Related Work

AL related approaches for ER include the common AL with the goal of selecting the most informative data for classifiers to be labeled by humans (single-model-based [18], committee-based [7,11,19,20,24,27]), and special AL approaches for the purpose of getting the best rules (like classifiers) that are able to provide high precision without considering the quality of training data [1–3,10,22]. Therein, Ngomo et al. [18] identify the most informative data to be labeled and added into the training dataset with the maximized convergence of the used classifier. The proposed committee-based AL approaches differ from each other globally with different committee forming approaches. The approaches [7,11,19,20] use genetic programming algorithms to learn multi-attribute functions. However, the quality of those functions cannot be guaranteed. The research approaches [24,27] are the most similar to ours. They form their committees with several classifiers, which are trained on a single type of classification algorithm. However, in order to achieve diversity of classifiers in the committee to make AL work with the most disagreement strategy, their classifier qualities are compromised, which restricts the ability of the committee to identify the most informative data. Moreover, the initial training dataset selection problem is not correctly handled. Sarawagi and Bhamidipaty [24] directly assume that the AL process starts with an initial training dataset including five matching and non-matching pairs, which is not realistic, since it cannot be known whether a pair is matching or non-matching before labeling. Although in the other paper [27], this reality is considered, however, the initial training dataset they selected is quite biased with the number of matching and non-matching pairs, which leads to a very low quality of classifiers for the beginning iterations. In contrast to them, our proposed HeALER can provide a high-qualified initial training dataset and the heterogenous committee can select more informative data to improve the classifiers faster.

5 Conclusions and Future Work

To conclude this paper, we propose our AL approach HeALER for ER, which could select relatively balanced and informative initial training dataset and use its heterogeneous committee to select informative pairs to be labeled by human in order to improve the classifier. We evaluated and compared it with the passive (machine) learning and two state-of-the-art AL-ER approaches ATLAS and ALIAS. The evaluation results show that HeALER is faster to converge and can reach a higher final F-measure than other approaches. In addition, the results also indicate that it requires less training data to reach a satisfactory F-measure, which conforms to the purpose of using AL approach: reducing human labeling effort. However, we also observed the fluctuations during the early rounds, which are caused by choosing outliers to the training dataset. For future work, techniques to exploit the local density to handle imbalanced data and recognize outliers [17] should be studied in order to improve HeALER and make it reach the convergence faster.

References

1. Arasu, A., Götz, M., Kaushik, R.: On active learning of record matching packages. In: SIGMOD, pp. 783–794 (2010)
2. Bellare, K., Iyengar, S., Parameswaran, A.G., Rastogi, V.: Active sampling for entity matching. In: SIGKDD, pp. 1131–1139 (2012)
3. Bellare, K., Iyengar, S., Parameswaran, A.G., Rastogi, V.: Active sampling for entity matching with guarantees. In: TKDD, pp. 12:1–12:24 (2013)
4. Chen, X., Durand, G.C., Zoun, R., Broneske, D., Li, Y., Saake, G.: The best of both worlds: combining hand-tuned and word-embedding-based similarity measures for entity resolution. In: BTW (2019)
5. Chen, X., Schallehn, E., Saake, G.: Cloud-scale entity resolution: current state and open challenges. In: OJBD, pp. 30–51 (2018)
6. Christen, P.: Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer Science & Business Media, Heidelberg (2012)
7. de Freitas, J., Pappa, G.L., da Silva, A.S., et al.: Active learning genetic programming for record deduplication. In: CEC, pp. 1–8 (2010)
8. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
9. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: a survey. In: IEEE TKDE, pp. 1–16 (2007)
10. Fisher, J., Christen, P., Wang, Q.: Active learning based entity resolution using Markov logic. In: Bailey, J., Khan, L., Washio, T., Dobbie, G., Huang, J.Z., Wang, R. (eds.) PAKDD 2016. LNCS (LNAI), vol. 9652, pp. 338–349. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-31750-2_27
11. Isele, R., Bizer, C.: Active learning of expressive linkage rules using genetic programming. J. Web Semant. **23**, 2–15 (2013)
12. Kotsiantis, S., Kanellopoulos, D., Pintelas, P., et al.: Handling imbalanced datasets: a review. GESTS Int'l. Trans. Comp. Sci. Eng. **30**(1), 25–36 (2006)
13. Leipzig, D.G.: Benchmark datasets for entity resolution (2017). Accessed 27 Nov 2017
14. Lu, Z., Wu, X., Bongard, J.: Active learning with adaptive heterogeneous ensembles. In: ICDM, pp. 327–336 (2009)
15. Mamitsuka, N.A.H., et al.: Query learning strategies using boosting and bagging. In: ICML (1998)
16. Melville, P., Mooney, R.J.: Diverse ensembles for active learning. In: ICML (2004)
17. Nanopoulos, A., Manolopoulos, Y., Theodoridis, Y.: An efficient and effective algorithm for density biased sampling. In: CIKM, pp. 398–404. ACM (2002)
18. Ngomo, A.N., Lehmann, J., Auer, S., Höffner, K.: RAVEN - active learning of link specifications. In: Proceedings of the International, Workshop on Ontology Matching (2011)

19. Ngonga Ngomo, A.-C., Lyko, K.: EAGLE: efficient active learning of link specifications using genetic programming. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 149–163. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-30284-8_17
20. Ngomo, A.-C.N., Lyko, K., Christen, V.: COALA – correlation-aware active learning of link specifications. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 442–456. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38288-8_30
21. Nguyen, H.T., Smeulders, A.: Active learning using pre-clustering. In: ICML, p. 79 (2004)
22. Qian, K., Popa, L., Sen, P.: Active learning for large-scale entity resolution. In: CIKM, pp. 1379–1388 (2017)
23. Rennie, J.D., Shih, L., Teevan, J., Karger, D.R.: Tackling the poor assumptions of Naive Bayes ext classifiers. In: ICML, pp. 616–623 (2003)
24. Sarawagi, S., Bhamidipaty, A.: Interactive deduplication using active learning. In: SIGKDD, pp. 269–278 (2002)
25. Seung, M.O., Sebastian, H., Sompolinsky, H.: Query by committee. In: Proceedings of the Workshop on Computational Learning Theory (1992)
26. Spark. Spark.mllib documentation. https://spark.apache.org/docs/latest/mllib-ensembles.html. Accessed 29 Nov 2018
27. Tejada, S., Knoblock, C.A., Minton, S.: Learning object identification rules for information integration. Inf. Syst. **26**, 607–633 (2001)
28. Wang, Q., Vatsalan, D., Christen, P.: Efficient interactive training selection for large-scale entity resolution. In: Cao, T., Lim, E.-P., Zhou, Z.-H., Ho, T.-B., Cheung, D., Motoda, H. (eds.) PAKDD 2015. LNCS (LNAI), vol. 9078, pp. 562–573. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18032-8_44