Active Gradual Machine Learning for Entity Resolution

Anonymous ACL submission

Abstract

Recent work has shown that the task of entity resolution (ER) can be effectively performed by gradual machine learning (GML). GML begins with some easy instances, which can be automatically labeled by the machine with high accuracy, and then gradually labels more challenging instances by iterative knowledge conveyance in a factor graph. Without involving manual labeling effort, the current GML solution for ER is unsupervised. However, its performance is limited by inaccurate and insufficient knowledge conveyance. Therefore, there is a need to investigate how to improve knowledge conveyance by manual labeling effort.

In this paper, we propose an active learning (AL) approach based on GML for ER. It iteratively generates new knowledge in the form of one-sided rules by manual label verification and instills them into a factor graph for improved knowledge conveyance. We first present a technique of knowledge discovery based on genetic mutations, which can generate effective knowledge rules with very small manual verification cost. Then, we demonstrate how to leverage the generated rules for improved knowledge conveyance by measuring their influence over label status by the metric of skyline distance. We have evaluated the performance of the proposed approach by a comparative study on real benchmark data. Our extensive experiments have shown that it can significantly improve the performance of unsupervised GML with very small manual cost; furthermore, it outperforms the state-of-the-art AL solutions for deep learning by considerable margins in terms of learning efficiency.

1 Introduction

Entity resolution (ER) aims at finding the records that refer to the same real-world entity (Barlaug and Gulla, 2021; Doan et al., 2020; Christen, 2012). Consider the running example shown in Table 1. ER needs to match the paper records between two tables, T_1 and T_2. The pair of $< e_{1i}, e_{2j} >$, in which e_{1i} and e_{2j} denote a record entity in T_1 and T_2 respectively, is called an *equivalent* pair if and only if e_{1i} and e_{2j} refer to the same paper; otherwise, it is called an *inequivalent* pair. In the example, e_{11} and e_{21} are *equivalent* while e_{12} and e_{22} are *inequivalent*.

The state-of-the-art solutions for ER were built on a variety of deep neural networks (DNN) (Li et al., 2020; Barlaug and Gulla, 2021; Mudgal et al., 2018; Ebraheem et al., 2018; Nie et al., 2019; Fu et al., 2019; Zhao and He, 2019)). However, to achieve high performance, they require a large quantity of accurately labeled training data, which unfortunately may not be readily available in real scenarios. Furthermore, DNN models usually have limited interpretability. To alleviate these limitations, a solution based on the paradigm of gradual machine learning (GML) has been recently proposed for ER (Hou et al., 2019; Hou et al., 2020). Without depending on the Independent and Identically Distributed (IID) assumption, GML begins with some easy instances, which can be automatically labeled by the machine with high accuracy, and then gradually reasons about the labels of more challenging instances by iterative knowledge conveyance in a factor graph. The current GML solution for ER does not require manual labeling effort, but its efficacy depends on effective knowledge conveyance from easy instances to harder ones. Unfortunately, unsupervised knowledge conveyance may be inaccurate and insufficient. On one hand, some pair instances may be mislabeled in the process of gradual learning, thus providing noisy evidential observations. On the other hand, the current solution conveys knowledge between instances by global influence regression based on pre-specified basic metrics, mostly value similarities on different attributes (e.g. paper titles or author names in the running example); however, learning effi-

Table 1: A running example of ER.

ID	Title	Author	Venue	Year
e_{11}	Peer Collaborative Learning for Online Knowledge Distillation	G. Wu, S. Gong	AAAI	2021
e_{12}	Deep Reinforcement Learning for General Game Playing	A. Goldwaser, M. Thielscher	AAAI	2020

T_1

ID	Title	Author	Venue	Year
e_{21}	Peer Collaborative Learning for Online Knowledge Distillation	Wu, Gong	AAAI	2021
e_{22}	Deep Reinforcement Learning for Navigation in AAA Video Games	Alonso, Peter, Goumard, Romoff	IJCAI	2021

T_2

ciency of such knowledge conveyance is limited because a handful of new observations could only have marginal impact on global distribution regression.

Therefore, there is a need to investigate how to enable supervised knowledge conveyance for improved gradual learning. Active learning (AL), in which data are actively sampled to be labeled by human oracles with the goal of maximizing model performance while minimizing labeling cost, has presented itself as a feasible approach for traditional machine learning (ML) models including DNN (Barlaug and Gulla, 2021; Doan et al., 2020; Settles, 2012). In this paper, we propose an active learning approach based on GML for ER. Instead of selecting samples for manual labeling and then submitting them for model training, the proposed approach leverage labeled samples to generate new knowledge in the form of one-sided rules and then instills them into GML factor graph for improved knowledge conveyance. Inspired by the concept of genetic evolution (Jong, 2006), it first generates a wide variety of candidate rules by mutations and then singles out the fittest among them by skyline observations with very small manual cost. The resulting rules can accurately indicate label status while covering many mislabeled instances. By measuring their influence over label status by skyline distance, the proposed approach enables effective knowledge conveyance with only a small amount of manual effort.

The major contributions of this paper can be summarized as follows:

1. We propose a novel active learning approach based on GML for ER, which can effectively improve the performance of gradual learning with only a small amount of manual effort;

2. We present a new technique of active knowledge generation for ER based on genetic evolution. It can generate highly accurate one-sided labeling rules based on skyline observations with very small manual cost;

3. We validate the efficacy of the proposed approach on real benchmark data by a comparative study. Our extensive experiments have shown that it can significantly improve the performance of GML with only a small amount of manual effort, and it considerably outperforms the state-of-the-art AL solutions for deep models in terms of learning efficiency.

2 Related Work

Due to space limit, we briefly review related work from the orthogonal perspectives of entity resolution and active learning.

Entity Resolution. The problem of ER has been extensively studied in the literature (Barlaug and Gulla, 2021; Doan et al., 2020; Christen, 2012). It has been widely recognized that the unsupervised approaches have limited efficacy in real scenarios (Bilenko et al., 2003). The supervised approaches viewed ER as a binary classification task and then applied various statistical learning models (e.g. SVM (Arasu et al., 2010; Bellare et al., 2012), native Bayesian (Berger, 1985), rule-based methods (Li et al., 2015; Quinlan, 1986) and DNN models (Mudgal et al., 2018; Li et al., 2020)) for the task. However, the performance of these supervised approaches heavily relies on labeled training data.

Recently, a non-i.i.d learning paradigm called *Gradual Machine Learning (GML)* (Hou et al., 2020; Hou et al., 2019; Zhong et al., 2021) has been proposed to enable effective machine learning for ER without the requirement for manual labeling effort. GML has also been applied to the task of sentiment analysis (Wang et al., 2021; Ahmed et al., 2021). The current unsupervised GML solutions can achieve competitive performance compared with many supervised approaches. However, without exploiting labeled training data, their performance is still limited by inaccurate and insufficient knowledge conveyance.

Active Learning. Active learning has been extensively studied in the context of machine learning. For traditional machine learning such as SVM, the most prominent approaches that proved to perform well include *margin-based*, *maximum entropy*, *Query by committee* and *Expected variance reduction* to name a few (Settles, 2012). However, many of the above methods pose challenges when applied to deep neural networks.

Most active learning works for DNN have been focused on image classification. They can be broadly categorized into three groups:(1) uncertainty-based (Houlsby et al., 2011; Gal and Ghahramani, 2016; Kirsch et al., 2019): they applied dropout at test time to approximate Bayesian inference enabling the application of Bayesian methods to deep learning; (2) expected model change-based (Zhang et al., 2017): they used an expected model change measure to choose examples that maximize the impact on the learned model weights when labeled; (3) representativeness-based (Ash et al., 2020; Yang et al., 2017; Elhamifar et al., 2013; Sener and Savarese, 2018): they usually aimed to achieve trade-off between representativeness and uncertainty. Other recent works include generative data augmentation for AL (Tran et al., 2019), e.g., adversarial network-based discrimination of informative points (Sinha et al., 2019) and detrimental point processes-based batch selection (Bıyık et al., 2019). Active deep learning for ER has also been specifically studied (Kasai et al., 2019; Bogatu et al., 2021). They usually tailored the mainstream AL strategies to ER.

3 Unsupervised GML for ER

Given an ER workload consisting of record pairs, a solution needs to label each pair in the workload as *equivalent* or *inequivalent*. The unsupervised GML solution for ER, as shown in Figure 1, consists of the following 3 essential steps:

3.1 Easy Instance Labeling.

Given an ER workload, unsupervised GML first uses an unsupervised clustering algorithm to estimate the proportions of equivalent and inequivalent instances in the workload, and then proportionally (e.g. 30%) identify the pair instances with the highest (resp. lowest) record similarities as the easy equivalent (resp. inequivalent) instances.

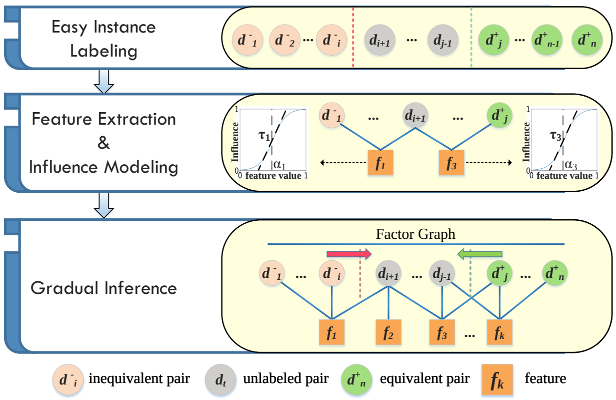


Figure 1: Overview of GML Solution for ER.

3.2 Feature Extraction and Influence Modeling.

GML extracts the features satisfying the monotonicity assumption of precision to facilitate knowledge conveyance, e.g. attribute value similarity and token features aligned with record similarity. Intuitively speaking, the monotonicity assumption of precision statistically states that an equivalence probability of a pair instance increases with its feature values. Since the proposed active learning approach also depends on the monotonicity assumption, we formally define it as in (Arasu et al., 2010):

Assumption 1 (Monotonicity of Precision) *A value interval I_i is dominated by another interval I_j, denoted by $I_i \preceq I_j$, if every value in I_i is less than every value in I_j. We say that precision is monotonic with respect to a pair metric if for any two value intervals $I_i \preceq I_j$ in [0,1], we have $P(I_i) \leq P(I_j)$, in which $P(I_i)$ denotes the equivalence precision of the set of instance pairs whose metric values are located in I_i.*

For each feature, GML models its influence over pair labels by a monotonous sigmoid function with two parameters, α and τ, which denote the function's midpoint and the steepness of the curve respectively. Formally, given a feature f and a pair d, the influence of f w.r.t d is represented by

$$P_f(d) = \frac{1}{1 + e^{-\tau_f (x_f(d) - \alpha_f)}}, \tag{1}$$

in which $x_f(d)$ represents d's feature value w.r.t f. According to Eq. 1, provided with the values of α_f and τ_f, the influence model statistically dictates that any feature value of $x_f(d)$ corresponds to an equivalence probability. Typically, the value of $P_f(d)$ increases with the feature value of d, or $x_f(d)$.

3.3 Gradual Inference.

GML fulfills gradual learning by a factor graph G, which consists of evidence variables Λ, inference variables V_I and factors modeling labeled instances, unlabeled instances and their shared features respectively. Typically, GML labels only one instance at each iteration. At each iteration, gradual inference essentially learns the feature parameter values (α and τ) such that the inferred results maximally match the evidential observations. Formally, the objective function can be represented by

$$(\hat{\alpha}, \hat{\tau}) = \arg\min_{\alpha,\tau} -\log \sum_{V_I} P_{\alpha,\tau}(\Lambda, V_I), \quad (2)$$

in which $P_{\alpha,\tau}(\Lambda, V_I)$ denotes the joint probability of the variables in G.

To enable scalable gradual learning, in each iteration, GML first selects the top-m unlabeled instances with the most evidential support as the candidates, and then efficiently approximates their probabilities. Finally, GML constructs factor graphs individually only for the top-k most promising unlabeled instances (or the instances with the lowest entropies) among the m candidates, to infer their probabilities via maximum likelihood. GML labels the one with the lowest entropy at each iteration. A newly labeled instance would serve as an evidential observation in the following iterations.

4 Active GML Framework

The active GML approach, denoted by A-GML, iteratively discovers new knowledge in the form of one-sided labeling rule and integrates them into GML factor graph for improved gradual learning. In each round, it can select some unlabeled instances from a pre-specified pool for manual label verification. As first introduced in (Chen et al., 2020), one-sided rules act as label status indicators. As opposed to the classical setting where a rule is used to label pairs in both ways (*equivalent* or *inequivalent*), a one-sided rule focuses exclusively on one single class. An example is as follows:

$$r_i[Year] \neq r_j[Year] \rightarrow inequivalent(r_i, r_j), \quad (3)$$

where $r_i[Year]$ denotes the record r_i's attribute value at *Year* and $inequivalent(r_i, r_j)$ denotes the inequivalence between r_i and r_j. With this knowledge, a pair of records with different publication years is supposed to have a high probability of being inequivalent. However, this rule does not intend to indicate the label status of any pair with the same publication year.

The active approach begins with the labeling result of unsupervised GML. As shown in Fig 2, given a manual budget of B for each round, it iteratively performs the following two steps: knowledge rule generation and rule-augmented gradual inference.

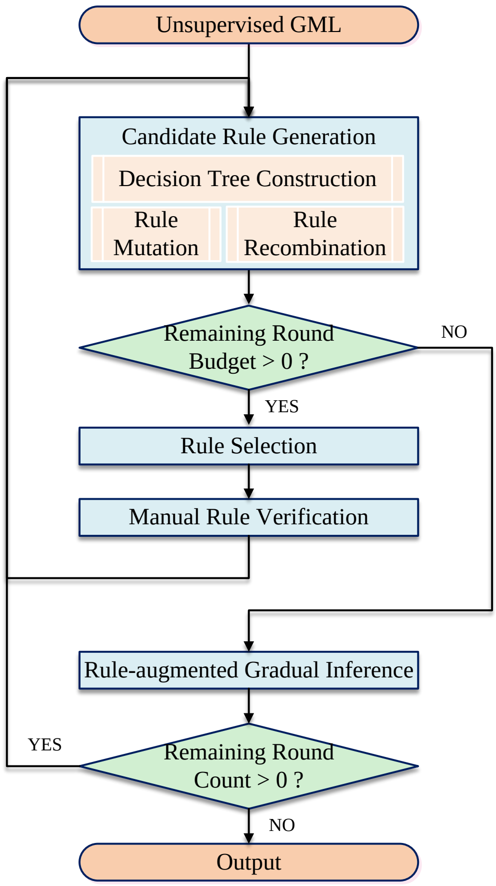


Figure 2: Active GML Framework for ER.

4.1 Knowledge Rule Generation

Formally, we represent a one-sided labeling rule by a first order logic expression as

$$\bigwedge_i p_i(f_i(d), v_i) \rightarrow L, \quad (4)$$

where $f_i(d)$ denotes the feature value of a pair instance d w.r.t a feature f_i, v_i denotes a constant within the domain of $f_i(d)$, $p_i(f_i(d), v_i)$ denotes a predicate in the form of $f_i(d) \leq v_i$ or $f_i(d) \geq v_i$, $\wedge$ denotes the conjunction of predicates, and L denotes a label status (0 for *inequivalent* or 1 for *equivalent*). To ensure high interpretability, we usually limit the maximum number of predicates in a rule to a small number (e.g. 2 in our implementation).

We discretize continuous feature values to facilitate knowledge generation. Specifically, given a feature, we deploy a Gaussian mixture model to cluster its values into a specified number of clusters (e.g. 20 in our experiments). We sort the ranks of clusters by their center values to preserve monotonicity. Then, given a continuous feature value, we assign the rank of its corresponding cluster as its discrete value.

The active approach generates an initial set of candidate rules based on the labeling result of unsupervised GML, and then tries to produce more candidates by genetic evolution. From the candidate rules, it singles out only a few fittest rules for manual label verification based on a reward metric. Intuitively speaking, only the highly accurate rules that can potentially correct many mislabeled instances are worthy of being verified. The detailed technical solution for knowledge rule generation will be presented in Section 5.

4.2 Rule-augmented Gradual Inference

Similar to unsupervised GML, A-GML also proportionally (e.g. 30% in our implementation) labels a set of easy equivalent and inequivalent instances based on record similarity to start gradual inference. In each round of A-GML, the manually verified instances are always considered as easy instances while the rest of easy instances are identified based on record similarity. For each validated rule, we create a corresponding factor, which are shared by all the satisfying instances. Similar to the existing GML features for ER, we also use the sigmoid function to model a rule's influence over label status as

$$w(f_r) = \tau \cdot (x_r - \alpha), \tag{5}$$

where r denotes a rule, f_r denotes the corresponding factor of r, and x_r denotes the feature value of an instance w.r.t f_r. With the monotonicity assumption, we quantify the feature value of x by an instance's skyline distance to the rule. The parameters of α and τ need to be learned based on evidential observations in the process of gradual inference. According to the sigmoid model, a rule's influence over an instance would increase with skyline distance. We will detail how to measure skyline distance in Section 5.

5 Knowledge Rule Generation

In this section, we describe how to efficiently generate accurate one-sided labeling rules with only a small amount of manual effort. The process consists of two steps: candidate rule generation and manual rule verification.

5.1 Candidate Rule Generation

Based on the labeling result of unsupervised GML, we directly use the algorithm proposed in (Chen et al., 2020) to generate the initial candidate rules with the maximal depth of m (m=2 in our experiments). To discover new knowledge beyond those implied by the initial rules, we explore new rules by the operations of gene mutation and gene recombination:

5.1.1 Gene Mutation

Consider the validated rule of

$$Sim(Title) \geq 18 \wedge Sim(Authors) \geq 18 \rightarrow 1, \tag{6}$$

which specifies that a pair instance is equivalent if both its discretized title similarity level and author similarity level are no less than 18. The mutation operation would relax the thresholds of title and author similarities by one discrete level. For instance, if both thresholds are relaxed to 17, we would get a new candidate rule represented as

$$Sim(Title) \geq 17 \wedge Sim(Authors) \geq 17 \rightarrow 1. \tag{7}$$

Due to the monotonicity assumption, we usually reduce the value level of equivalence predicates while increasing the value level of inequivalence predicates. Since the number of predicates in any rule is small, our algorithm executes all possible relaxation operations on a validated rule to generate as many candidate rules as possible.

5.1.2 Gene Recombination

Beside gene mutations, we also extract the predicates (genes) from separate rules and combine them by the AND operator to reproduce new ones. It is noteworthy that the recombination operation has to be executed on the predicates indicating the same label. Since the total number of defined predicate templates is limited (e.g. only dozens in our experiments) and the maximum number of predicates in a rule is small (e.g. 2 in our implementation), we construct all possible predicate combinations, whose total number is also limited.

5.2 Manual Rule Verification

Rules are supposed to be verified based on skyline observations. Therefore, in this subsection, we first introduce the concept of skyline distance, and then describe how to efficiently select accurate rules with potential big reward with only a small amount of manual cost.

5.2.1 Skyline Distance

Given a set of instances of D and a rule of r, an instance d_i in D is said to be a skyline of a rule r if and only if d_i is not strictly dominated by any other instance in D w.r.t r. Given an equivalence rule of r, an instance d_i is said to strictly dominate another one d_j, $r: d_i \succ d_j$, if and only if d_i's value at each predicate of r is no larger than that of d_j and there exists at least one predicate such that d_i's value is less than that of d_j. It is noteworthy that according to the monotonicity assumption, if d_i strictly dominates d_j, the equivalence probability of d_j is at least as large as that of d_i. The case for inequivalence rule is similar.

Building upon the work in (Huang et al., 2013), we define a non-skyline instance's skyline distance to a rule as follows:

Definition 1 ***Skyline Distance.*** *Given a rule, r, the skyline distance of an instance $d_i \in D$ to r, denoted by $SkyDist_r(d_i)$, is defined as the minimum sum of the changing values on all the predicates of r to move d_i to a new position d_i', so that d_i' is not strictly dominated by any other instance in D. That is, $SkyDist_r(d_i) := min_{d_i',\, r:\, d_i' \succeq d_i,\, \nexists d_j \in D,\, r:\, d_j \succ d_i'} MD(d_i, d_i')$, where $MD(d_i, d_i')$ denotes the Manhattan Distance between d_i and d_i'.*

Our strategy of rule verification is built upon the monotonicity assumption of skyline distance, which can be formally stated as follows:

Assumption 2 (Monotonicity Assumption of Skyline Distance) Given a rule of r indicating the label of L (L=0 or 1), an interval of skyline distance I_i is dominated by another interval I_j, denoted by $I_i \preceq I_j$, if every skyline distance in I_i is no less than every skyline distance in I_j. We say that precision is monotonic with respect to a skyline distance if for any two skyline distance intervals $I_i \preceq I_j$ in [0,1], we have $P(I_i) \geq P(I_j)$, in which $P(I_i)$ denotes the precision that the labels of the set of instances whose skyline distance values are located in I_i are equivalent to L.

5.2.2 Rule Selection

In each round, we iteratively select the rule with the maximum reward for manual verification until the round budget of B runs out. Formally, the reward of a rule r, $W(r)$, is estimated by

$$W(r) = Conf(r) \cdot Benf(r), \tag{8}$$

where $Conf(r)$ represents the confidence of r, and $Benf(r)$ denotes the benefit of r, or the number of instances whose currently predicted labels is not consistent with r (can thus be potentially corrected by r). Specifically, we measure $Conf(r)$ by the difference between the estimated equivalence probabilities of r's skylines, S_r, and their labels as indicated by r:

$$Conf(r) = |1 - L - \frac{1}{|S_r|} \sum_{d_i \in S_r} P(d_i)|. \tag{9}$$

In Eq. 9, if d_i has a ground-truth label, its value of $P(d_i)$ is equal to 0 (if inequivalent) or 1 (if equivalent). If d_i does not have a ground-truth label, its value of $P(d_i)$ is approximated by the equivalence probability estimated by the current GML model.

After manual verification, if the proportion of r's skyline observations, whose ground-truth labels match r's indicating label, exceeds a pre-specified threshold θ (e.g. θ=0.95 in our implementation), the rule is considered to be true and will participate in the next round of gradual inference. In case that a chosen candidate rule fails manual verification, the algorithm would try to produce new candidate rules by re-constructing one-sided decision trees based on new manual observations as well as the current GML labeling results.

5.2.3 Discussion on Verification Efficiency

Rule verification generally requires to manually inspect every skyline. It is noteworthy that for ER, the maximum number of predicates in rules needs to be limited to a small value (e.g., 2 in our implementation) to ensure high interpretability. As a result, the number of a rule's skylines is usually small (e.g., dozens in our experiments) in most cases. Furthermore, we set a budget of B° (B°=20 in our implementation) for each rule's verification. If a rule has less than B° skylines, those with the smallest skyline distances would be additionally verified. In case that the number of skylines exceeds B°, the algorithm would select the instances in the decreasing order of entropy as predicted by the current GML model.

5.3 An Illustrative Example

We illustrate the process of rule generation by the examples extracted from the DBLP-ACM workload[1]. Active GML generates a candidate rule based on the results of unsupervised GML in the first round as follows:

$$r_1 : Eq(Year) = 0 \wedge Sim(Authors) \leq 16 \rightarrow 0, \quad (10)$$

where the predicate of *Eq(Year)* indicates whether two records have the same publication year, and *Sim(Authors)* denotes their value similarity at *Authors*. By gene mutation, the following new candidate rule is generated in the second round:

$$r_2 : Eq(Year) = 0 \wedge Sim(Authors) \leq 17 \rightarrow 0. \quad (11)$$

After r_2 is verified to be valid, the threshold of *Sim(Authors)* continues to be relaxed. Finally, the discrete level of *Sim(Authors)* reaches the maximum and r_1 evolves into the following rule with only one predicate:

$$r_3 : Eq(Year) = 0 \rightarrow 0. \quad (12)$$

On DA, the results of unsupervised GML contain many false positives, many of which however can be successfully predicted by r_3. As a result, the first round of active GML can significantly improve precision as shown in our empirical evaluation.

6 Empirical Evaluation

In this section, we empirically evaluate the performance of the proposed approach (denoted by **A-GML**). Besides against the unsupervised GML solution, we have compared A-GML with four state-of-the-art deep AL solutions tailored to the Ditto (Li et al., 2020), which is the state-of-the-art deep model for ER. The four AL solutions include: 1) Maximum Entropy (Yang and Loog, 2018) (denoted by **ME-Ditto**). The traditional approach samples the points with the highest entropy values in each round; 2) BALD (Houlsby et al., 2011) (denoted by **BALD-Ditto**). Also based on uncertainty measurement, it samples the points that maximize the mutual information with Ditto's parameters; 3) EGL (Zhang et al., 2017) (denoted by **EGL-Ditto**). Based on the metric of expected model change, it samples the points that cause the biggest change to the embedding layer parameters of DNN; 4) BADGE (Ash et al., 2020) (denoted by **BADGE-Ditto**). The recently proposed approach samples points with diverse gradient embeddings to trade off between uncertainty and diversity.

The evaluation has been conducted on four widely used benchmark datasets, which include: 1) Abt-Buy[1] (denoted by AB): ER needs to match product entities from two commercial websites, Abt.com and Buy.com; 2) DBLP-ACM[1] (denoted by DA): ER needs to match the publication entities from two sources, DBLP and ACM; 3) Songs[2] (denoted by SG): ER needs to match the song entries within a single table; 4) iTunes-Amazon[1] (denoted by IA): ER needs to match the music entities from two sources, iTunes and Amazon.

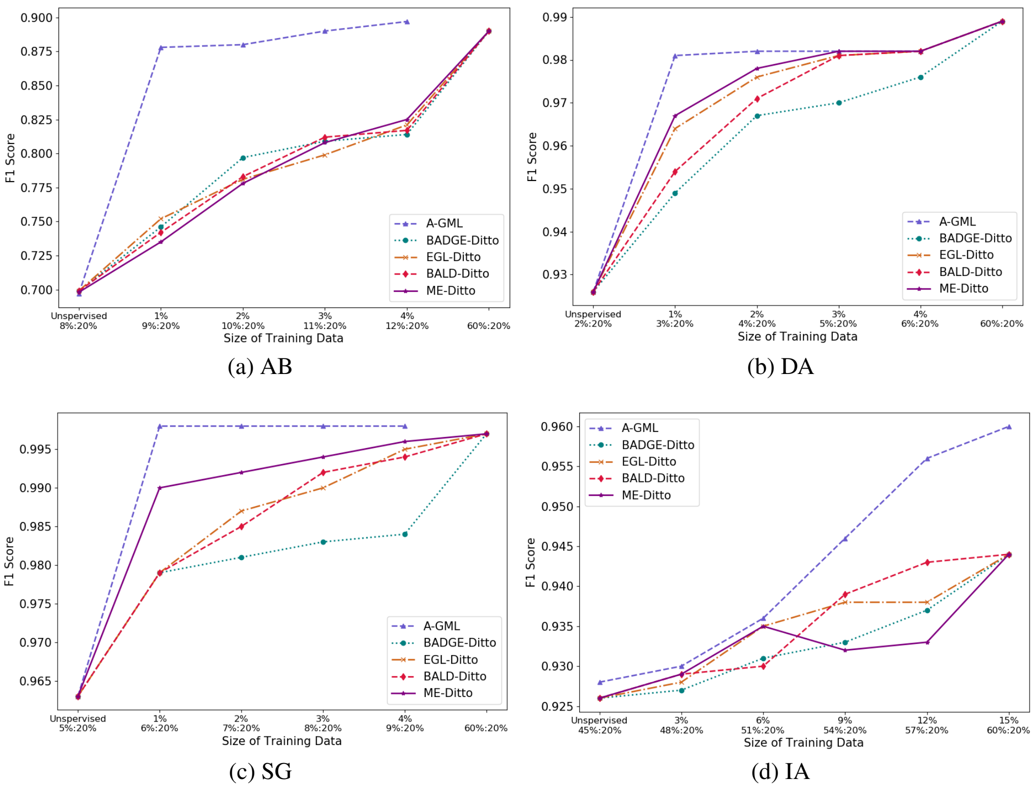
As in (Li et al., 2020), we randomly divide each dataset into three parts, the training pool (60%), the validation pool (20%) and the test pool (20%). Active instances are sampled from the training pool while performance is evaluated on the test pool. The validation pool is used for Ditto's hyperparameter tuning. Note that unsupervised GML can achieve competitive performance compared with supervised Ditto. For fair comparison, in the evaluation of AL solutions for Ditto, we randomly select an initial set of instances from the training pool to train Ditto such that its performance is very close to that of unsupervised GML. With the similar initial performance, we then compare A-GML and various AL solutions for Ditto in terms of learning efficiency. On AB, DA and SG, each AL round samples 1% instances from the training pool for manual verification; while on IA, each round samples 3% due to its small data size.

As usual, we measure performance by the metric of F1, which is a balanced combination of precision and recall. In the implementation of A-GML, we set the maximum size of verified skyline set at 20. We set the discrete levels of continuous metric values at 20. The accuracy threshold of a valid rule is set at 95%. The performance of A-GML is observed to be very stable. The performance of Ditto is however relatively more volatile. All the reported results are averages over 5 runs. The codes are available at https://github.com/wailler/ActiveGML.

Evaluation Results: the detailed evaluation results

[1]available at https://github.com/megagonlabs/ditto

[2]available at http://pages.cs.wisc.edu/~anhai/data/falcon_data/songs



(a) AB (b) DA

(c) SG (d) IA

Figure 3: Comparative Evaluation Results.

have been presented in Fig. 3. It can be observed that the supervised approach improves the performance of GML by considerable margins. On most datasets (e.g. AB, DA and SG), the performance of GML almost reaches the maximum with only two rounds, but flattens out thereafter. The only exception is on IA, where the performance of A-GML can consistently improve in the five rounds. It is worthy to point out that this observation should not be surprising, because the performance of A-GML is destined to be theoretically bounded by the expressive power of one-sided rules. Since all the rules are constructed based on the existing basic metrics, there exist some instances in each dataset that no rule is able to correctly label without compromising overall labeling quality.

It can be observed that with the similar initial performance, A-GML performs considerably better than the deep AL solutions in terms of learning efficiency on all the test datasets. Specifically, with only one round (manual cost at 1%), A-GML achieves the close-to-optimal performance on AB, DA and SG while the deep AL solutions take considerably more rounds. We also report the optimal performance that can be achieved by Ditto provided with all the labeled data in training pools. With 60% of the whole dataset as training data, Ditto can be supposed to be sufficiently trained. It can be observed that on AB and SG, A-GML, which exploits only 4% training data, beats the Ditto model trained with 60% training data. On IA, A-GML provided with only 20% training data also beats Ditto trained with 60%. On DA, Ditto trained with 60% however beats A-GML trained with 4% with a slight margin. Our experimental results clearly demonstrate the efficacy of A-GML.

7 Conclusion

In this paper, we have proposed a novel active learning approach based on GML for ER. By generating accurate one-sided labeling rules based on skyline observations, it can effectively improve the performance of GML with very small manual cost. Our empirical study has validated its efficacy. For future work, we have observed that not surprisingly, the performance of the active solution is limited by the expressive power of rules constructed based on pre-specified basic metrics; unfortunately, increasing the number of predicates in a rule has limited efficacy. Therefore, it is interesting to investigate other forms of knowledge that can further improve the performance of GML in future work.

References

Murtadha Ahmed, Qun Chen, Yanyan Wang, Youcef Nafa, Zhanhuai Li, and Tianyi Duan. 2021. DNN-driven gradual machine learning for aspect-term sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 488–497, Online. Association for Computational Linguistics.

Arvind Arasu, Michaela Götz, and Raghav Kaushik. 2010. On active learning of record matching packages. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2010, Indianapolis, Indiana, USA, June 6-10, 2010*, pages 783–794.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Nils Barlaug and Jon Atle Gulla. 2021. Neural networks for entity matching: A survey. *ACM Trans. Knowl. Discov. Data*, 15(3).

Kedar Bellare, Suresh Iyengar, Aditya G. Parameswaran, and Vibhor Rastogi. 2012. Active sampling for entity matching. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, pages 1131–1139.

James O. Berger. 1985. *Statistical Decision Theory and Bayesian Analysis, 2nd Edition*. Springer.

Mikhail Bilenko, Raymond J. Mooney, William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5):16–23.

Erdem Bıyık, Kenneth Wang, Nima Anari, and Dorsa Sadigh. 2019. Batch Active Learning Using Determinantal Point Processes. *arXiv e-prints*.

Alex Bogatu, Norman W. Paton, Mark Douthwaite, Stuart Davie, and André Freitas. 2021. Cost-effective variational active entity resolution. In *37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021*, pages 1272–1283. IEEE.

Zhaoqiang Chen, Qun Chen, Boyi Hou, Zhanhuai Li, and Guoliang Li. 2020. Towards interpretable and learnable risk analysis for entity resolution. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, SIGMOD '20, page 1165–1180, New York, NY, USA. Association for Computing Machinery.

Peter Christen. 2012. *Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications. Springer.

AnHai Doan, Pradap Konda, Paul Suganthan G. C., Yash Govind, Derek Paulsen, Kaushik Chandrasekhar, Philip Martinkus, and Matthew Christie. 2020. Magellan: Toward building ecosystems of entity matching solutions. *Communications of the ACM*, 63(8):83–91.

Muhammad Ebraheem, Saravanan Thirumuruganathan, Shafiq Joty, Mourad Ouzzani, and Nan Tang. 2018. Distributed representations of tuples for entity resolution. *Proc. VLDB Endow.*, 11(11):1454–1467.

E. Elhamifar, G. Sapiro, A. Yang, and S. S. Sasrty. 2013. A convex optimization framework for active learning. In *2013 IEEE International Conference on Computer Vision*, pages 209–216.

Cheng Fu, Xianpei Han, Le Sun, Bo Chen, Wei Zhang, Suhui Wu, and Hao Kong. 2019. End-to-end multi-perspective matching for entity resolution. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4961–4967. International Joint Conferences on Artificial Intelligence Organization.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1050–1059.

B. Hou, Q. Chen, Y. Wang, Y. Nafa, and Z. Li. 2020. Gradual machine learning for entity resolution. *IEEE Transactions on Knowledge and Data Engineering*.

Boyi Hou, Qun Chen, Jiquan Shen, Xin Liu, Ping Zhong, Yanyan Wang, Zhaoqiang Chen, and Zhanhuai Li. 2019. Gradual machine learning for entity resolution. In *Proceedings of the 2019 The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 3526–3530.

Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *CoRR*, abs/1112.5745.

Jin Huang, Bin Jiang, Jian Pei, Jian Chen, and Yong Tang. 2013. Skyline distance: a measure of multidimensional competence. *Knowl. Inf. Syst.*, 34(2):373–396.

Kenneth A. De Jong. 2006. *Evolutionary computation - a unified approach*. MIT Press.

Jungo Kasai, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa. 2019. Low-resource deep entity resolution with transfer and active learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5851–5861, Florence, Italy. Association for Computational Linguistics.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc,

E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7026–7037. Curran Associates, Inc.

Lingli Li, Jianzhong Li, and Hong Gao. 2015. Rule-based method for entity resolution. *IEEE Trans. Knowl. Data Eng.*, 27(1):250–263.

Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. 2020. Deep entity matching with pre-trained language models. *Proc. VLDB Endow.*, 14(1):50–60.

Sidharth Mudgal, Han Li, Theodoros Rekatsinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra. 2018. Deep learning for entity matching: A design space exploration. In *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*, pages 19–34.

Hao Nie, Xianpei Han, Ben He, Le Sun, Bo Chen, Wei Zhang, Suhui Wu, and Hao Kong. 2019. Deep sequence-to-sequence entity matching for heterogeneous entity resolution. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 629–638. Association for Computing Machinery.

J. Ross Quinlan. 1986. Induction of decision trees. *Machine learning*, 1(1):81–106.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Burr Settles. 2012. *Active Learning*. Morgan & Claypool Publishers.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. 2019. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Toan Tran, Thanh-Toan Do, Ian Reid, and Gustavo Carneiro. 2019. Bayesian generative active deep learning. volume 97 of *Proceedings of Machine Learning Research*, pages 6295–6304. PMLR.

Yanyan Wang, Qun Chen, Jiquan Shen, Boyi Hou, Murtadha Ahmed, and Zhanhuai Li. 2021. Aspect-level sentiment analysis based on gradual machine learning. *Knowledge-Based Systems*, 212:106509.

Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z. Chen. 2017. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*.

Yazhou Yang and Marco Loog. 2018. A benchmark and comparison of active learning for logistic regression. *Pattern Recognit.*, 83:401–415.

Ye Zhang, Matthew Lease, and Byron C. Wallace. 2017. Active discriminative text representation learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3386–3392. AAAI Press.

Chen Zhao and Yeye He. 2019. Auto-em: End-to-end fuzzy entity-matching using pre-trained deep models and transfer learning. In *The World Wide Web Conference*, WWW '19, page 2413–2424. Association for Computing Machinery.

Ping Zhong, Zhanhuai Li, Qun Chen, and Boyi Hou. 2021. Attention-enhanced gradual machine learning for entity resolution. *IEEE Intelligent Systems*.