Explainable Data Matching: Selecting Representative Pairs with Active Learning Pair-Selection Strategies

Lukas Laskowski, Florian Sold[1]

Abstract: In both research and enterprise, dirty data poses numerous challenges. Many data cleaning pipelines include a data deduplication step that detects and removes entries within a given dataset which refer to the same real-world entity. Throughout the development of such deduplication techniques, data scientists have to make sense of the large result sets that their matching solutions generate to quickly identify changes in behavior or to discover opportunities for improvements. We propose an approach that aims to select a small subset of pairs from the result set of a data matching solution which is representative of the matching solution's overall behavior. To evaluate our approach, we show that the performance of a matching solution trained on pairs selected according to our strategy outperforms a randomly selected subset of pairs.

Keywords: Entity Resolution; Data Matching; ExplainableDM; Pair Selection; Benchmark

1 Explainable Data Matching

Improving data matching systems is an iterative process: Insights on matching behavior derived from the set of output labels of the matching solution serve as the basis for improvements in the next iteration. To accelerate this optimization process, we have developed a data matching benchmark platform, called Frost [Gr22]. Frost combines existing benchmarks, established quality metrics, cost and effort measures for evaluating and comparing data matching solutions. Furthermore, the platform includes techniques which enable the systematic exploration of matching results. However, as real-world datasets can contain millions of records, it is unrealistic to examine all pairs within a result set. Consequently, there is the need to summarize data matching results such that only a representative subset of the most meaningful pairs remains. Based upon a matching result set of size m generated with a data matcher on a dataset of size n, we aim to select a subset of well-distinguishable pairs of size k with $k \ll m$ that are representative of a matching solution's behavior.

To achieve this goal, we leverage instance selection strategies from the field of active learning, a semi-supervised learning method, where the initial seeded training dataset is very small or empty. Therefore, to sufficiently train the data matching classifier, more label

[1] Hasso Plattner Institute, University of Potsdam, Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, {lukas.laskowski, florian.sold}@student.hpi.de

doi:10.18420/BTW2023-77

candidates are iteratively selected and, after a manual labeling process, added to the training dataset. Since manual labeling is costly, a variety of pair selection approaches exist that aim to achieve a target matching quality in as few iterations as possible. Our goal to select a subset of pairs which summarize the matching classifier as good as possible is similar to the pair-selection task in active learning. Nevertheless, we differ in how the pair-selection algorithm is applied since our use case works upon a fully labelled result set and therefore lacks the iterative procedure.

Within the field of active learning for data matching, Christen et al. [CCR20] recently proposed an iterative pair-selection strategy based upon their novel informativeness measure, which outperforms prior works. Hence, we base our non-iterative pair-selection approach upon the informativeness measure by Christen et al. With an informativeness score for each pair, we choose the representative subset as the set of pairs with the highest score.

In the following sections, we first introduce related work to explainable data matching (Section 2) and then introduce our proposed approach (Section 3). Next, in Section 4, we show the effectiveness of this strategy by experimentally analyzing its behavior. Finally, we conclude and discuss next steps (Section 5).

2 Related Work

Explaining data matching decisions is generally a challenging task. Especially with complex machine learning or deep learning models like DITTO [Li21], results are difficult to interpret but often superior to those achieved with approaches based upon simple classifiers or rule-based approaches. Without specific tools, it is close to impossible to understand how a certain black-box matcher assigns labels, as the underlying model can be very complex. Additionally, result sets are large and make it difficult to select pair instances for further analysis. While explanation techniques already exist in the machine learning community, Thirumuruganathan et al. found that these techniques do not suit the needs of data matching scientists and propose a variety of research opportunities [TOT19].

Baraldi et al. [Ba21] propose, with their Landmark explanation framework, a new framework for local pair-specific explanations specifically tailored towards data matching. Landmark introduces two main innovations: First, it generates per pair two explanations by fixing either record (called landmark) and perturbing the other record (called varying entity). Second, it produces an artificial entity by attribute-wise concatenating both entities. The MOJITO framework [Di19] analyzes the influence of individual attributes on the matching decision. Their results show that black-box entity matching models might rely on untrustworthy attributes. Hence, they conclude that quality metrics are not sufficient to quantify real-world performance of a data matching model, but rather emphasize the need for explanations.

Nevertheless, surprisingly little work has been done so far towards matching solution-agnostic filtering or selection of representative (or summarized) result subsets. Explanation frameworks like Landmark can then be executed upon pairs within a representative subset.

3 Informativeness-based Selection of Pairs

To explain the behavior of a matcher, we present an approach that selects pairs out of the result set labeled by a matching solution. As described in Section 1, we base our selection strategy upon the informativeness measure developed by Christen et al. [CCR20].

$$informativeness(u, R) = (1 - \alpha) * entropy(u, R) + \alpha * uncertainty(u, R)$$

The informativeness-score for u in respect to the result set R is the weighted average of the uncertainty and the entropy. The balancing factor is given as α and set to $\alpha = 0.5$.

To each labeled pair within R, we assign a similarity vector u, which consists of the pair-wise similarity for each attribute of the two records r_1, r_2 with $(r_1, r_2) \in R$. To calculate the informativeness, we use an additional similarity measure between a pair of pair vectors u, v, in our case cosine similarity, called $sim(u, v)$ that we assume to already exist. For a pair u, we define the set R_S as all pairs from the result set that were assigned the same label as u and as R_O those assigned the opposite label. An environment S of supporting pairs for a pair u is bounded by the closest instance classified contrary to the target pair, and defined as:

$$S(u, R) = \{v \in R_S | sim(u, v) > max\{sim(u, w) | w \in R_O\}\}$$

$$uncertainty(u, R) = \frac{1}{1 + |R \cap S(u, R)|}$$

Uncertainty quantifies how certain a label assignment is. In case only very few similar pairs reside within the environment S around a target pair, the label assigned to it is uncertain. Hence, its uncertainty score will be higher. Compared to a pair with plenty of confirmations, a pair with high uncertainty might be especially valuable and representative of the matching classifier's behavior.

$$entropy(u, R) = -[\frac{\sum_{v \in R_{SE}} u * sim(u, v)}{|R| - 1} * log(\frac{\sum_{v \in R_{SE}} u * sim(u, v)}{|R| - 1}) + \frac{\sum_{v \in R_{OE}} u * sim(u, v)}{|R|} * log(\frac{\sum_{v \in R_{OE}} u * sim(u, v)}{|R|})]$$

The entropy value represents how diverse the environment around a certain pair u is. It can only yield a high value in case both summands are high negative values. Again, this can only be true in case we have both very similar pairs of the same and the opposite class. Pairs that fulfill this requirement are considered to have high entropy, as they help to form the decision boundary and are therefore, again, representative of the classifier's behavior. Compared to the active learning setting, our approach does not work iteratively, and we do have already all pair labels available (as assigned by the matching solution). Therefore, we would need to consider all pairs for entropy calculation, which in return would lead to a very similar entropy value for all pairs. Hence, we restrict the pairs considered for entropy calculation to those pairs whose similarity to the target pair u is higher than the entropy environment limit e. We define the environment boundary e relative to the S-Environment

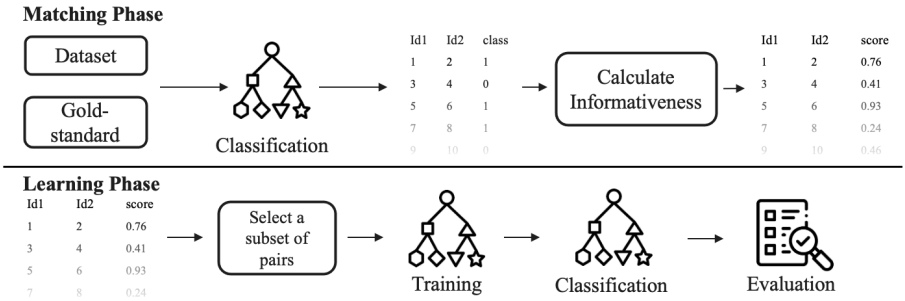

Fig. 1: **Evaluation Process.** This figure shows the two phases of our evaluation approach: "Matching Phase" and "Learning Phase".

as $e(u, R) = S(u, R) * l$ with $l \in [0, 1]$ and set to $l = 0.4$. The set R_{SE} includes all pairs of the same class as u and a similarity larger than $e(u, R)$ to u. Similarly, we define the set R_{OE} with pairs of opposite class. Since the entropy base metric requires a similarity value for all pairs of pairs, the proposed approach has a runtime complexity of $O(n^4)$ with n as the amount of records in the base dataset.

Under the assumption that machine learning models benefit similarly from a representative pair subsets as humans do, we select the top k pairs by informativeness measure as the representative subset. Hence, this particular subset summarizes the matching behavior of the matching solution better than any other equally sized subset could. In case one chooses k small enough, the selected pairs now serve as a basis for a human analysis.

4 Evaluation

We evaluate our approach by training a matching solution only upon the ground truth labels of pairs within the representative subset, and then comparing its quality using F1-Score against two baselines. Figure 1 outlines the overall evaluation-process: We first produce results for informativeness-calculation by predicting labels on unseen testing data, which serve as the basis for the informativeness score ("Matching Phase"). During the subsequent "Learning Phase", we select the k pairs with the highest informativeness score as our subset of pairs. These pairs and their respective ground truth label are then used to train the same (but untrained) classification model. Afterwards, we predict and evaluate against the full gold standard.

We perform experiments using the matching solution "Cyber-Punk" which is one of the winning concepts of the SIGMOD programming contest in 2021[2]. As the dataset, we used "SIGMOD AltoSight Z4", which contains mainly textual data about SD cards. Besides our informativeness selection strategy, we compare up with two naïve baseline subsets:

[2] https://dbgroup.ing.unimore.it/sigmod21contest/index.shtml

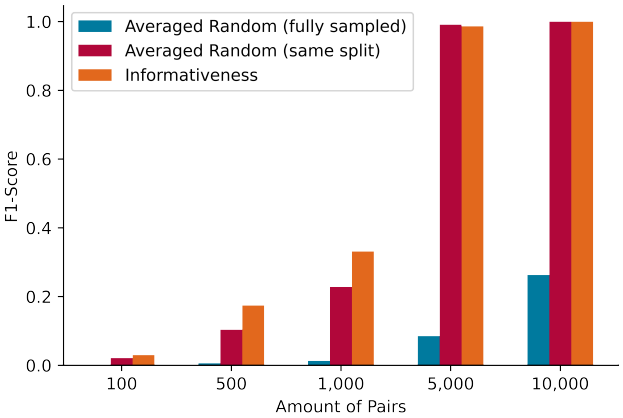

Fig. 2: **Retraining on Selected Pairs.** Comparing the results for the presented selection strategies on multiple subset sizes on dataset *Z4* using the "Cyber-Punk" matcher. We use subset sizes of $k \in \{100, 500, 1000, 5000, 10000\}$.

- **Averaged Random (fully sampled)**: For this strategy, we sample randomly k different pairs out of the set of pairs. In practice, this resembles a human who scrolls through the entire result set with no further filters or selections available.
- **Averaged Random (same split)**: In most datasets, the majority of pairs are easy to label as non-matches. Therefore, the previous strategy predominantly selects such pairs and only few (if any) duplicate pairs. In contrast, this strategy samples k pairs with the same proportion of duplicates and non-duplicates (according to the ground truth annotation) as can be found in the selected subset. Although we sample randomly, this subset implicitly comes with a substantial advantage as its proportions are based upon the informativeness-based subset with respect to the ground truth annotation.

As shown in Figure 2 the informativeness-based selection strategy indeed outperforms the completely randomly sampled selection strategy at any subset size. This observation is reasonable, as likely many trivial non-matches were sampled into the subset. Consequently, a smaller subset size k causes a larger difference between the two solutions: At a subset size of 100, the informativeness-based strategy ($f_1 = 0.03$) has a 15x higher F1-Score compared to the randomly based selection strategy ($f_1 = 0.002$). With more pairs in the subset, we see a similar effect. For instance, the informativeness-based strategy ($f_1 = 0.986$) has a subset size of 5,000 an F1-Score which is 11.6x higher than the randomly based selection strategy ($f_1 = 0.085$). Even in comparison to the random sampling of *same split*, our approach achieves the highest scores on small subsets and matches on subsets of size $k \geq 5000$.

Since our use-case is mainly targeted towards subsets of size 1000 or less that a human can

grasp or look through, and we do significantly outperform both sampling approaches in this area, our approach does indeed work as anticipated. This shows that a subset selected using the informativeness score includes pairs that bear more information compared to random sampling – and therefore offers insights into the behavior of a matching solution.

5 Conclusion and Outlook

We set out to select a representative subset of pairs out of a potentially very large result set. Our results do indeed indicate that the informativeness-based subset outperforms random selection by far. More importantly, these results now serve as a baseline for further improvements towards reducing the overall runtime or developing novel approaches.

Beyond our current results, we see further research opportunities in this area. For example, our existent evaluation can be underlined with additional test settings including multiple datasets as well as a diverse set of matching solutions from various domains. Furthermore, one could further extend this approach by indicating for each selected pair how many similar non-selected pairs exist in the result set. In case these pairs were accessible to a data scientist, he could better make sense of this particular pair's properties.

Acknowledgements. This paper is the result of a seminar supervised by Felix Naumann and Luca Zecchini. We thank them for their valuable input and support during our project.

References

[Ba21] Baraldi, Andrea; Del Buono, Francesco; Paganelli, Matteo; Guerra, Francesco: Landmark Explanation: An Explainer for Entity Matching Models. In: Proceedings of the International Conference on Information and Knowledge Management. ACM, pp. 4680 – 4684, 2021.

[CCR20] Christen, Victor; Christen, Peter; Rahm, Erhard: Informativeness-Based Active Learning for Entity Resolution. In: European Conference on Principles of Data Mining and Knowledge Discovery. Springer International Publishing, pp. 125–141, 2020.

[Di19] Di Cicco, Vincenzo; Firmani, Donatella; Koudas, Nick; Merialdo, Paolo; Srivastava, Divesh: Interpreting Deep Learning Models for Entity Resolution: An Experience Report Using LIME. In: Proceedings of the Second International Workshop on Exploiting Artificial Intelligence Techniques for Data Management. ACM, 2019.

[Gr22] Graf, Martin; Laskowski, Lukas; Papsdorf, Florian; Sold, Florian; Gremmelspacher, Roland; Naumann, Felix; Panse, Fabian: Frost: a platform for benchmarking and exploring data matching results. PVLDB, 15(12):3292–3305, 2022.

[Li21] Li, Yuliang; Li, Jinfeng; Suhara, Yoshihiko; Doan, AnHai; Tan, Wang-Chiew: Deep Entity Matching with Pre-Trained Language Models. PVLDB, 14(01):50–60, 2021.

[TOT19] Thirumuruganathan, Saravanan; Ouzzani, Mourad; Tang, Nan: Explaining Entity Resolution Predictions: Where are we and What needs to be done? In: Proceedings of the Workshop on Human-In-the-Loop Data Analytics. ACM, pp. 1–6, 2019.