Active Learning of Equivalence Relations by Minimizing the Expected Loss Using Constraint Inference

Steffen Rendle, Lars Schmidt-Thieme
Machine Learning Lab
Institute for Computer Science
University of Hildesheim, Germany
{srendle, schmidt-thieme}@ismll.uni-hildesheim.de

Abstract

Selecting promising queries is the key to effective active learning. In this paper, we investigate selection techniques for the task of learning an equivalence relation where the queries are about pairs of objects. As the target relation satisfies the axioms of transitivity, from one queried pair additional constraints can be inferred. We derive both the upper and lower bound on the number of queries needed to converge to the optimal solution. Besides restricting the set of possible solutions, constraints can be used as training data for learning a similarity measure. For selecting queries that result in a large number of meaningful constraints, we present an approximative optimal selection technique that greedily minimizes the expected loss in each round of active learning. This technique makes use of inference of expected constraints. Besides the theoretical results, an extensive evaluation for the application of record linkage shows empirically that the proposed selection method leads to both interesting and a high number of constraints.

1 Introduction

Several important tasks in the area of machine learning can be formalized as finding an equivalence relation. Clustering and record linkage are two prominent examples. In this paper we investigate equivalence relations in general, i.e. we do not make any assumptions about the number of equivalence classes or about class sizes in the target relation. Throughout the paper we illustrate learning an equivalence relation by the application of record linkage. There are many applications for record linkage like merging stocks from different e-commerce websites or finding identical people in social networks. Estimating the target equivalence relation is done by learning a similarity measure using training data [6, 10]. For acquiring training data, active learning can be used that chooses the most promising pairs [10, 11, 2] and presents it to the supervisor. The quality of a predicted relation mainly depends on the learned similarity measure. Thus the task of active learning is to create a good training set that generalizes well to unseen pairs. Furthermore the queried pairs can be used as constraints on the target relation to restrict the set of possible solutions [7].

The scope of this work are selection techniques for equivalence relations. One central point of this paper is, that by querying the right pairs of objects, additional constraints can be inferred to enlarge the training set. This is motivated by the fact, that the pairs have to form an equivalence relation. Thus, in the first part we will examine the properties of constraints and derive theoretical upper and lower bounds on the number of queries needed to converge to the optimal solution. In the second part, we investigate optimal loss reduction. Here we assume that a similarity function is learned on the set of constraints. We derive a selection criterion that greedily minimizes the expected loss between the learned similarity function and the expected equivalence relation. Finally, we show how our selection technique can be applied to the task of record linkage. In our evaluation we compare our method of expected loss minimization to other selection techniques of the record linkage and the semi-supervised-clustering community.

2 Related Work

Active learning is well studied in the field of classification. Popular approaches are reducing the size of the version space [3] or selecting by maximal uncertainty [5]. Roy and McCallum [9] have suggested a method for minimizing the expected error for classification. There are several important differences between classification and predicting equivalence relations which leads to other optimal selections.

Active learning for record linkage aka object identifica-

tion, duplicate detection, etc. has already been studied by several researchers [10, 11, 2]. Tejada et al. [11] as well as Sarawagi and Bhamidipaty [10] suggest to query the pair that is most uncertain for a committee of classifiers. They investigate in depth the setup of the committee, i.e. which classifiers should be chosen. Bilenko and Mooney [2] propose a method for selecting interesting pairs to build a training set. Their selection technique is combining the selection of the most similar pair and selecting a random pair. In total the proposed selection techniques for record linkage are sampling by uncertainty [10, 11] and sampling by a combination of most similar and random pairs [2].

There is also research on selection techniques for active learning in the related community of semi-supervised clustering. For problems with an unknown number of classes, Basu et al. [1] propose the EXPLORE algorithm that tries to find as much different clusters as possible. In our evaluation we will show that this approach fails for problem settings with a high number of classes as mostly cannot-links are found. Huang et al. [4] select pairs that score highest with regard to a gain model where the gain depends on the current clustering assignment. Their method assumes that the number of clusters is known in advance.

3 Learning Equivalence Relations

3.1 Constraint Inference

A common way for representing structural knowledge on a clustering problem X is to use sets of must-link and cannot-link constraints over pairs of objects. Let $D \subseteq X^2$ be the set of all given constraints. Let $D_m := \{(x, y) \in D | x \equiv y\}$ be the set of must-links in D and let $D_c := \{(x, y) \in D | x \not\equiv y\} = D \setminus D_m$ be the set of cannot-links in D. Two objects that are constrained by a must-link are equivalent objects, i.e. they belong to the same equivalence class. A cannot-link constraint over two objects indicates that the two objects are different.

Lemma 3.1 (Constraint inference) *The set of all inferable constraints $\overline{D} \supseteq D$ from a given set D is the closure under (1) and (2) where (1) are the equivalence axioms:*

$$\begin{aligned} &(x, x) \in \overline{D}_m && \textit{reflexive} \\ &(x, y) \in \overline{D}_m \rightarrow (y, x) \in \overline{D}_m && \textit{symmetric} \\ &(x, y) \in \overline{D}_m \wedge (y, z) \in \overline{D}_m \rightarrow (x, z) \in \overline{D}_m && \textit{transitive} \end{aligned} \quad (1)$$

(2) are the inferable cannot-links:

$$(x, y) \in \overline{D}_m \wedge (x, z) \in D_c \rightarrow (y, z) \in \overline{D}_c \quad (2)$$

With $\overline{D}_m := \left(\overline{D}\right)_m$ and $\overline{D}_c := \left(\overline{D}\right)_c$.

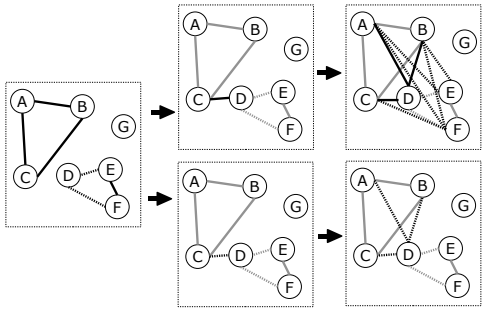

Figure 1. Constraint inference: Querying the pair (C,D) will either result in a must-link-constraint (top) or a cannot-link-constraint (bottom). Additional constraints can be inferred (right). Must-links are bold lines; cannot-link dashed ones.

The reason for this is, that first the elements in D_m have to form an equivalence relation, i.e. have to meet the axioms in (1). Secondly, $\overline{D}_m$ has to be the smallest equivalence relation over must-link constraints in D. We will use the common notation $[x]_E := \{y | (x, y) \in E\}$ for describing the equivalence class of x for an equivalence relation E and we write $x \equiv_E y$ iff $(x, y) \in E$. Finally, one can also infer new cannot-link constraints by combining given must-links and cannot-links by using formula (2).

Let D^* be the training set extended with a new labeled pair (x^*, y^*): $D^* := D \cup \{(x^*, y^*)\}$ We will write D^+ if we assume that $x^* \equiv y^*$ otherwise D^-. An example for constraint inference can be found in Figure 1, where the system queries the pair (C, D).

3.2 Solutions

A solution for a record linkage problem $X = \{x_1, \ldots, x_n\}$ is an equivalence relation E over X^2. Given pairwise must-link and cannot-link constraints $D = D_m \cup D_c$ the set of consistent solutions is:

$$\mathcal{E}_D = \{E \subseteq X^2 | E \supseteq \overline{D}_m \wedge E \cap \overline{D}_c = \emptyset\}$$

We will also use the term *version space* for the set of consistent solutions $\mathcal{E}_D$.

3.3 Active Learning for Equivalence Relations

The objective of an active learning algorithm is to converge with as little effort for the supervisor to the optimal solution E^*. In each round of active learning, the algorithm selects a pair of objects $(x, y) \in X^2$ that should be labeled

by the supervisor as identical $x \equiv y$ or different $x \not\equiv y$. In our work we assume a faultless oracle that never fails.

Afterwards, normally [10, 11, 2] the labeled pair is used to improve the learned similarity measure. In our work, we also investigate the constraints that are induced by the pair. That means, (x, y) is added to the must-links if the two objects are labeled to be equivalent $x \equiv_{E^*} y$ – otherwise if $x \not\equiv_{E^*} y$ then (x, y) is added as a cannot-link. The new constraint will reduce the version space.

k rounds of active learning can be formalized as choosing object pairs $(x_1, y_1), \ldots, (x_k, y_k)$ and obtain boolean labels $(l_1, \ldots, l_k)$ from the supervisor. Active learning with constraints infers $\overline{D}^1, \ldots, \overline{D}^k$ that results in version spaces $\mathcal{E}_{\overline{D}^1}, \ldots \mathcal{E}_{\overline{D}^k}$. If only non-trivial pairs (i.e. $X^2 \setminus \overline{D}$) are queried, then the number of constraints increases strictly monotonic and thus also the size of the version space decreases strictly monotonic $|\mathcal{E}_{\overline{D}^i}| > |\mathcal{E}_{\overline{D}^{i+1}}|$ until convergence to E^*.

3.4 Bounds on the Number of Queries

Next we derive tight upper and lower bounds on the number of queries that are necessary to restrict the set of possible solutions $\mathcal{E}$ of a problem X to a single solution E^*.

Lemma 3.2 (Bounds on the number of queries) *The number $|D|$ of queries obeys the following bounds:*

$$|X| + \frac{1}{2}k \cdot (k-3) \leq |D|$$
$$\leq \frac{1}{2}|X| \cdot (|X|+1) - k - \sum_{i=1}^{k} \frac{|c_i| \cdot (|c_i|-1)}{2}$$

Where the perfect solution E^* consists of the classes $c_1, \ldots, c_k$. Both bounds are tight.

For the lower bound, at least one cannot-link constraint between two instances of each class is necessary to separate them. Additionally in each class there has to be a chain of must-link constraints that links all objects inside the class. The exact upper bound on the number of constraints is given by the worst selection technique that first selects all non-equivalent pairs and then the equivalent pairs.

To give a practical example, the 112 class Cora dataset with 1295 objects would need at least 7399 queries to converge to a single solution. The worst selection technique would present 821,864 pairs before converging.

4 Optimal loss reduction for active learning

First we show how reduction of expected loss can be used in a general classification setting. Afterwards we derive a method for the task of learning equivalence relations.

4.1 Loss reduction for classification

The general classification task is to assign a class $c \in C$ to each object $x \in X$. Roy and McCallum [9] define the classification error for the finite data set S and a labeled set D as follows:

$$E_{\hat{P}_D} = \frac{1}{|S|} \sum_{x \in S} L(P(c|x), \hat{P}_D(c|x)) \tag{3}$$

Where L is an arbitrary loss function and $P(c|x)$ is the true but unknown class assignment of an instance x.

The task of active learning is to converge to the best quality (lowest error) with as few effort for the supervisor as possible. An optimal greedy selection, selects in each round the object x that will result in the minimal possible error. We suggest to use averaging over the class probabilities which corresponds to calculating the expected value:

$$\operatorname*{argmin}_{x \in X} \sum_{c \in C} P(c|x) E_{\hat{P}_{D \cup \{(x,c)\}}} \tag{4}$$

Our formulation differs from the approach of Roy and McCallum who instead use an optimistic estimate of c and minimize the following objective:

$$\operatorname*{argmin}_{x \in X} \min_{c \in C} E_{\hat{P}_{D \cup \{(x,c)\}}}$$

As P is never observed, we can estimate it by $\hat{P}_{D \cup \{x,c\}}$. With this estimate of P our minimization criterion (4) corresponds to minimizing the expected error.

4.2 Loss reduction for equivalence relations

Now we derive an optimal loss minimization for equivalence relations. As we will see, the inference of constraints plays an important role in the minimization.

For equivalence relations, the items are pairs of objects $(x, y) \in X^2$ and the class is binary, stating the equivalence of a pair, i.e. $x \equiv y$ or $x \not\equiv y$. With this binary relation over X^2 the pool-based error (3) can be reformulated as:

$$E_{\hat{P}_D} = \frac{1}{|X^2|} \sum_{(x,y) \in X^2} L(P(x \equiv y), \hat{P}_D(x \equiv y))$$

For L a loss function like 0-1 loss (L_0) or log-loss (L_l) can be used:

$$L_0(p, \hat{p}) := p \cdot \delta(\hat{p} > 0.5) + (1-p) \cdot \delta(\hat{p} \leq 0.5)$$
$$L_l(p, \hat{p}) := -p \operatorname{ld} \hat{p} - (1-p) \operatorname{ld}(1-\hat{p})$$

Here δ is an indicator function that evaluates the truth of a statement:

$$\delta(s) := \begin{cases} 1, & \text{if } s \text{ is true} \\ 0, & \text{otherwise} \end{cases}$$

4.2.1 Constraints on the target relation

The probability $\hat{P}_D$ is estimated not only from D but also from set of all inferable constraints $\overline{D}$ (see section 3.1). For unseen data, the estimate $\hat{P}_D$ can be generated by an arbitrary probabilistic classifier. To summarize, the estimated probability $\hat{P}_D$ is defined as:

$$\hat{P}_D(x \equiv y) := \begin{cases} 0, & (x,y) \in \overline{D}_c \\ 1, & (x,y) \in \overline{D}_m \\ \hat{P}'_D(x \equiv y), & \text{otherwise} \end{cases}$$

Where $\hat{P}'_D$ is the pairwise similarity estimated by a classifier.

With this definition, the error (3) can be simplified to[1]:

$$E_{\hat{P}_D} = \frac{1}{|X^2|} \sum_{(x,y) \in X^2 \setminus \overline{D}} L(P(x \equiv y), \hat{P}'_D(x \equiv y))$$

4.2.2 Approximating the optimal selection

In an active learning setting, we present the supervisor the pair (x^*, y^*) that will result in the minimum expected error. With the notation of section 3.1 the optimization criterion (4) can be rewritten as:

$$\underset{(x^*,y^*) \in X^2}{\operatorname{argmin}} \left(P(x^* \equiv y^*) E_{\hat{P}_{D^+}} + P(x^* \not\equiv y^*) E_{\hat{P}_{D^-}} \right) \tag{5}$$

Lemma 4.1 (Optimal selection) *If one approximates $\hat{P}'_{D^*}$ with $\hat{P}'_D$ the optimal query w.r.t. (5) and a given loss function L is:*

$$\begin{aligned} &\underset{\substack{(x^*,y^*) \in X^2, \\ (x^*,y^*) \notin \overline{D}}}{\operatorname{argmax}} \sum_{(x,y) \in \overline{D^-} \setminus \overline{D}} L(P(x \equiv y), \hat{P}'_D(x \equiv y)) \\ &+ P(x^* \equiv y^*) \sum_{(x,y) \in \overline{D^+} \setminus \overline{D^-}} L(P(x \equiv y), \hat{P}'_D(x \equiv y)) \end{aligned} \tag{6}$$

The first part sums the loss over all pairs between the two equivalence classes $[x]_{\overline{D}_m}$ and $[y]_{\overline{D}_m}$. The second part sums the loss between all pairs additionally inferred by (2) under the assumption that x is equivalent to y.

For performing the optimal selection in formula (6) a further improvement can be made. Instead of searching the best pair (x^*, y^*) on an object level (i.e. $X^2 \setminus C(D)$), the search can be performed on the level of equivalence classes. Thus for each pair $([x^*]_{\overline{D}_m}, [y^*]_{\overline{D}_m})$ of equivalence classes induced by the current constraint set D the optimization criterion has only to be computed once. In this case, the probability $P(x^* \equiv y^*)$ outside the sums of formula (6) can be estimated by the mean of the similarities between $[x^*]_{\overline{D}_m}$ and $[y^*]_{\overline{D}_m}$:

$$P\left([x]_{\overline{D}_m} \equiv [y]_{\overline{D}_m}\right) := \underset{\substack{x' \in [x]_{\overline{D}_m}, \\ y' \in [y]_{\overline{D}_m}}}{\operatorname{avg}} P(x' \equiv y')$$

As P is unobserved, similar to Roy and McCallum we suggest to approximate it with the current estimate $\hat{P}_D$. In total, two approximations have been performed for optimizing formula (5):

1. P is approximated by $\hat{P}_D$. This approximation is done as P is unobserved.
2. $\hat{P}'_{D^*}$ is approximated by $\hat{P}'_D$. This is done to simplify the optimization criterion and to prevent retraining the classifier for each unlabeled pair, which would require retraining the model $O(X^2 \setminus \overline{D})$ times.

5 Experiments

To evaluate the proposed selection methods, we apply them to a (semi-)supervised problem of record linkage [6, 10, 7].

5.1 Evaluation Scenarios

There are basically two application scenarios for active learning for equivalence relations:

1. **Optimal overall quality**: A problem X is given and one is interested in the optimal solution on X. E.g. a company has some domain experts that should deduplicate one of their databases. The goal is to get maximal quality on X with a minimal amount of queries.
2. **Generalization task**: One could also be interested in the generalization ability of the learned similarity measure on a further dataset. E.g. a model for a certain domain should be learned on some parts of X and after the active learning phase is applied to new data X_2.

The difference between the 'optimal quality' and the 'generalization task' is that in the optimal quality task one can ask questions on the same dataset where the evaluation takes place.

5.2 Datasets and Model Setup

We evaluate on the bibliographic *Cora* dataset[2] and on the product dataset *Camera* of a price comparison system[3]. As both datasets contain many trivial identical records,

[1] Assuming that the loss is 0 for correct predictions.

[2] http://www.cs.cmu.edu/~wcohen/match.tar.gz

[3] Mentasys GmbH, Germany, http://www.mentasys.de/

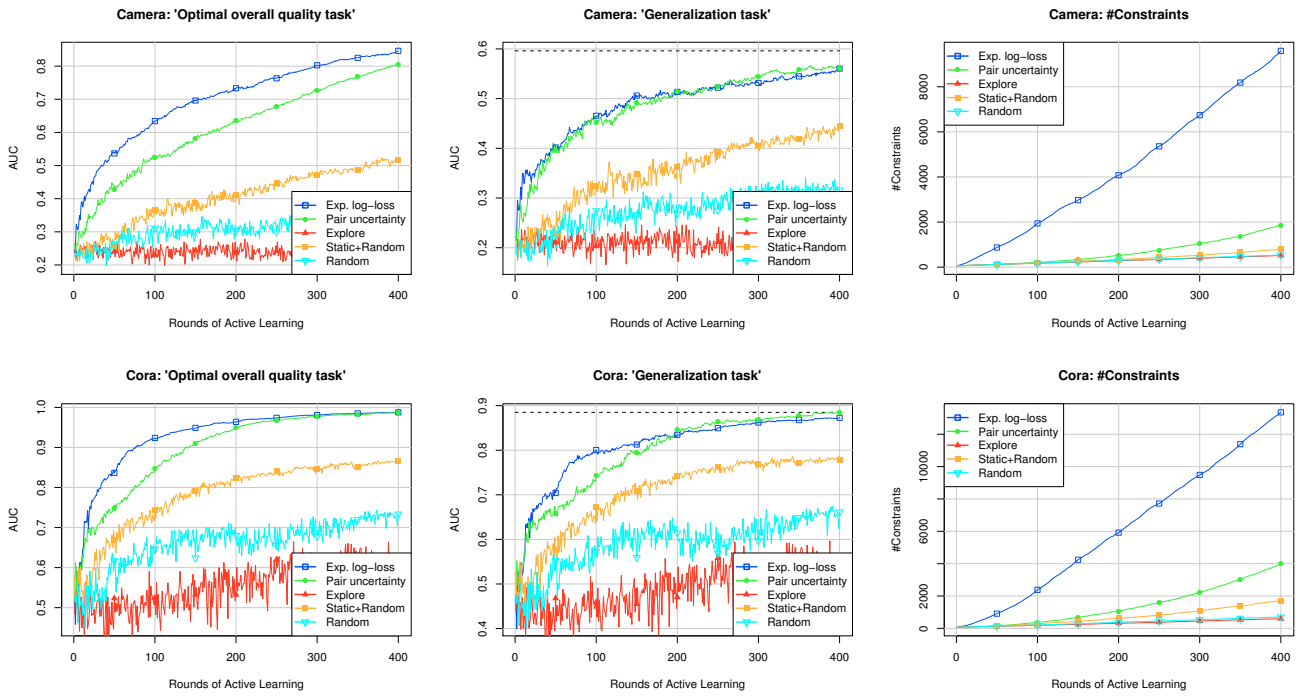

Figure 2. AUC performance and number of queried+inferred constraints of the proposed EXPECTED LOSS **technique in comparison to** PAIR UNCERTAINITY **similar to Sarawagi and Bhamidipaty 2002 [10] and Tejada et al. 2002 [11],** EXPLORE **of Basu et al. 2004 [1],** STATIC + RANDOM **of Bilenko and Mooney 2003 [2] and the baseline** RANDOM**.**

we generated reduced datasets by performing the reduction steps proposed by Rendle and Schmidt-Thieme [8]. The reduction results in two smaller subsets containing all challenging objects where Cora consists of 655 objects and Camera of 12722 objects.

The models for *Cora* and *Camera* are set up as follows. We use the same features and blocker as in [8]. For learning the similarity measure P'_D we use stacking with logistic regression as meta-classifier and as base-classifiers we use an ADTree, a J48 tree and logistic regression from Weka[4] and a C-SVM from libSVM[5]. The features for the meta-classifier consist of the probabilistic estimates of the four base-classifiers. When learning the similarity measure from a set of constraints D, we first split $\overline{D}$ in 10 folds and generate cases for the meta-classifier by training each base-model on 9 folds and predicting the remaining part. This is repeated for all 10 folds and results in training cases for the meta-model on the complete set $\overline{D}$. After training the meta-model on the generated cases, each base-classifier is retrained on $\overline{D}$.

[4] http://www.cs.waikato.ac.nz/ml/weka/
[5] www.csie.ntu.edu.tw/~cjlin/libsvm/

5.3 Experimental Methodology

First, we randomly divide the data set X in two parts of equal size: one part X_{sel} where selections can be performed and one part X_{gen} where the generalization is measured. We start our experiments with randomly picking 25 equivalent and 25 non-equivalent pairs from X_{sel}, $|D_m| = 25$ and $|D_c| = 25$. This is done to have enough training data to initialize the learned similarity measure. Now active learning starts and in each round the model selects one object pair $(x^*, y^*) \in X^2_{\text{sel}} \setminus \overline{D}$ with its selection technique. This pair is then labeled by a faultless oracle and added to the training set D^*. New constraints are inferred from the labeled pair and existing pairs using the rules (1) and (2). Then, the labeled data including inferred constraints is used to retrain the similarity measure.

In each round of active learning, we measure the quality of the similarity measure and the number of constraints. As quality measure we use the area-under precision-recall-curve (AUC) on pair level with respect to the learned similarity measure $\hat{P}_D$. The quality is both measured on X_{sel} and X_{gen}. The quality on X_{sel} measures the success in terms of the 'Optimal overall quality' task and the quality on X_{gen}

gives information about the generalization capabilities of the learned model.

We repeat all experiments 10 times and report the mean. For Cora we use different initializations of the random seed generator and run the experiments on all 655 items. For Camera we run the experiments on ten non-overlapping subsets with each about 1272 items. In total our evaluation is quite extensive and took about 90 days of CPU time.

5.4 Results and Discussion

Figure 2 on the right shows the number of all constraints depending on the rounds of active learning. As one can see, our selection technique EXPECTED LOSS leads to much more constraints than any other technique. Figure 2 also shows the AUC of both evaluation scenarios. First of all, one can see that EXPECTED LOSS and PAIR UNCERTAINTY outperform the other selection techniques on both datasets and evaluation schemes. For the task of minimizing the overall error, EXPECTED LOSS clearly outperforms PAIR UNCERTAINTY on both datasets. For example, after 100 rounds of active learning EXPECTED LOSS reaches an AUC-score of 0.64 on the Camera dataset whereas PAIR UNCERTAINTY has only 0.54 – likewise after 200 rounds EXPECTED LOSS reaches 0.74 and outperforms PAIR UNCERTAINTY with 0.64. On the much easier Cora dataset, the difference in the first 200 rounds is again large, before both techniques converge to an AUC of almost 1.0. For the task of generalization, one can see that both EXPECTED LOSS and PAIR UNCERTAINITY converge to the equivalence relation that could be achieved when training on the whole set X^2_{sel} (see dashed line in figure 2). On Camera the generalization capabilities of both methods are almost the same, whereas on Cora EXPECTED LOSS outperforms PAIR UNCERTAINITY clearly on the first 100 rounds. Another interesting result is that EXPLORE [1] is very inefficient for both record linkage tasks and is even outperformed by RANDOM. The reason for this is that mostly cannot-links are selected.

6 Conclusion

In this paper we have investigated active learning for predicting equivalence relations. First we have analyzed the problem setting and the implications of constraints. We have shown that by selecting certain pairs additional constraints can be inferred to enlarge the training data. We have derived both upper and lower bounds on the number of queries needed to converge to a single solution. The main contribution of this paper is the presented approximative optimal selection technique for minimizing the expected loss. In our evaluation we have shown that our proposed selection technique outperforms other state-of-the-art techniques in terms of finding many constraints and prediction quality.

Acknowledgements

This work was funded by the X-Media project (www.x-media-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-026978.

References

[1] S. Basu, A. Banerjee, and R. J. Mooney. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the SIAM International Conference on Data Mining (SDM-2004)*, 2004.

[2] M. Bilenko and R. J. Mooney. On evaluation and training-set construction for duplicate detection. In *Proceedings of the 2003 ACM SIGKDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, pages 7–12, 2003.

[3] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[4] R. Huang, W. Lam, and Z. Zhang. Active learning of constraints for semi-supervised text clustering. In *Proceedings of the Seventh SIAM International Conference on Data Mining (SDM-2007)*, 2007.

[5] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12, 1994.

[6] A. K. McCallum, K. Nigam, and L. Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the 6th International Conference On Knowledge Discovery and Data Mining (KDD-2000)*, pages 169–178, Boston, MA, Aug. 2000.

[7] S. Rendle and L. Schmidt-Thieme. Object identification with constraints. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM-2006)*, Hong Kong, 2006.

[8] S. Rendle and L. Schmidt-Thieme. Scaling record linkage to non-uniform distributed class sizes. In *In Proceedings of the 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2008)*, Osaka, 2008.

[9] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 441–448, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[10] S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*, pages 269–278, Edmonton, Alberta, 2002.

[11] S. Tejada, C. A. Knoblock, and S. Minton. Learning domain-independent string transformation weights for high accuracy object identification. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*, pages 350–359, Edmonton, Alberta, 2002.