Learning Balls from Correction Queries

Leonor Becerra-Bonache LEONOR.BECERRA.BONACHE@UNIV-ST-ETIENNE.FR
Colin de la Higuera CDLH@UNIV-ST-ETIENNE.FR
Jean Christophe Janodet JANODET@UNIV-ST-ETIENNE.FR
Frederic Tantini FREDERIC.TANTINI@UNIV-ST-ETIENNE.FR
Laboratoire Hubert Curien (ex-EURISE), 18 rue du Professeur Benoît Lauras, 42000 Saint-Etienne, France

1. Introduction

Learning in a noisy setting is a very hard task within the field of Grammatical Inference, even if the problem has been addressed for a long time now (de la Higuera, 2005). In this work, we will try to deal with this problem, by learning balls from correction queries.

A reasonable way to model noisy data is through the edit distance. But for this distance, language classes from the standard Chomsky hierarchy prove to be quite inadequate. Taking into account these two points, we will try to learn topological balls in the context of query learning.

D. Angluin introduced in (Angluin, 1987) the query learning model, which allows the learner to make queries to a teacher about the target language. Although membership queries (MQs) and equivalence queries (EQs) have established themselves as the standard combination to be used, there are real grounds to believe that EQs are too powerful to exist or even be simulated. Based on the growing evidence that corrections are available to children, we propose to work with another kind of query called correction query (CQ). This idea was introduced for the first time in (Becerra-Bonache & Yokomori, 2004).

We consider that CQs can play an important role in a noisy context. Thanks to a CQ, data that do not belong to the target concept are corrected.

Therefore, in this paper we will explore the relevance of CQs within the Grammatical Inference framework. Since EQs are computationally costly and without correspondence in a real life setting, we will base our study on learning from only CQs or MQs. In that way, we will try to answer the following question: "are CQs more powerful than MQs?"

2. Preliminaries

The *edit distance* between two strings is given by the minimum number of operations needed to transform one string into the other, where an operation is an *insertion*, *deletion*, or *substitution* of a single character.

In this work we consider a teacher able of answering *correction queries*. For an arbitrary string w submitted by the learner, if $w \notin L$, then the teacher returns a *correcting string of* w *with respect to* L. The correction of w will be a string c in L such that $d(w, c)$ is minimum, where $d(\cdot, \cdot)$ is the edit distance and each weight of the three operations is equal 1. If there is more than one possible correction, arbitrarily one of them will be returned. In case $w \in L$, the teacher's answer is "yes".

We use the edit distance to define the class of balls $\mathcal{B}_\Sigma$: the *ball* of center $u \in \Sigma^*$ and radius $r \in \mathbb{N}$ is $B_r(u) = \{w \in \Sigma^* | d(w, u) \leq r\}$. The set of strings with maximum length in the ball is denoted by B_{max}. Note that strings of B_{max} are obtained from u using only r insertions. Hence, all the letters of u appear in these strings, with the right order. Moreover, let $\Sigma = \{a_1, \ldots, a_n\}$, the words of type $a_i^r u$ are the unique words in B_{max} such that a_i^r is on the left of the center. These words will play an important role in order to infer the ball.

3. Learning with Correction Queries

We will study the complexity of learning balls using MQs or CQs, and we will compare the obtained results. In that way we try to prove that CQs are more powerful than MQs.

Since the answer of a MQs is only "yes" or "no", learning balls from only MQs requires an exponential number of queries: $O(\exp(|\Sigma| + |u| + r))$. Suppose we want to learn a ball which contains only one word of length n. The learner starts asking λ and then, he enumerates the words of length $1, 2, \dots, n$. The answer to all these MQs will be *no*, until he asks the correct string to the teacher. Clearly, he will need an exponential number of MQs in order to infer the ball.

This result led us to the following question: what about learning balls from MQs but giving to the learner one string that belongs to the ball? We present an algorithm that requires only a polynomial number of MQs when one positive data is provided to the learner: $O(|\Sigma| \cdot (|u| + r))$. The main idea of this algorithm is: first, from the positive data, find a string of B_{max} by inserting symbols to the current string and asking to the teacher whether the string obtained belongs to the ball or not; second, from this string of B_{max}, get $a_1^r u$ and $a_n^r u$, which correspond to the shortest and longest string (in alphabetic order) of B_{max}, by using swapping operations; finally, comparing these two strings we are able to know the center and radius of the ball.

Therefore, taking into account these two results, we try to learn balls using only CQs and see the difference with respect to learning from MQs.

Note that the same algorithm used for learning balls from "MQs and one positive data" could be also applied to learn balls from only CQs, but in this case, we would be able to get one element of the ball by asking for λ (we would obtain one of the shortest strings of the ball). The problem of this approach is that the remainder of the algorithm do not take into account the corrections received and only MQs would be enough.

Since we are interested in the properties of CQs, we present another algorithm which takes into account the corrections received during the process. In that way we can also see if correction queries truly enable us to learn balls in a more efficient way. Such learning algorithm consists of: first, find strings with the maximum number of each letter of the alphabet and then, thanks to this information, obtain one string of B_{max}; second, from the last string, look for one string of the form $a_i^r u$ using swapping operations. For the first step, only a logarithmic number of CQs is required ($O(|\Sigma| + \log(|u| + r))$); in that moment, we can know the radius and we can have also information about the center. However, for the last step, a linear number of queries is required in order to identify the center ($O(|\Sigma| + |u| + r)$), which increases the total complexity of the algorithm.

Comparing this result with the previous ones, we can see that there is a significant difference between learning balls from MQs or CQs only. However, when one positive data is available to the learner this difference is not so big from a theoretical stand point. That is why it would be interesting to learn balls using a logarithmic number of CQs.

Then, this leads us to the following question: can we get better results using another kind of correction? In order to answer this question, we introduce here another kind of correction based on *weighted edit distance*, which assigns different cost to the three basic operations. In that way, we base our corrections assuming the following weights: $substitution = 1 - \varepsilon$, $insertion = 1$, $deletion = 1$.

Using this kind of correction we obtain a curious result. If $|\Sigma| = 2$, it is still required a linear number of CQs (we could not avoid swapping operations in order to get the center). Though, if $|\Sigma| \neq 2$ we need only a logarithmic number of CQs: $O(|\Sigma| + \log(|u| + r))$.

4. Conclusions and Further Work

The results obtained show that CQs are more powerful than MQs. Although in some cases the difference between learning balls from CQs and MQs is not clear (using the standard edit distance), in other cases the difference is significant (using the weighted edit distance and $|\Sigma| \neq 2$).

As a future work, we will try to learn balls using a logarithmic number of CQs in the case $|\Sigma| = 2$, and using also the standard edit distance. In order to do that, we have to solve the problem of finding the center of the ball without using swapping operations (this is what increases the complexity of our algorithm). We would also like to do some experiments in order to see the difference about the number of queries used in practice.

References

Angluin, D. (1987). Learning regular sets from queries and counterexamples. *Information and Computation, 75*, 87–106.

Becerra-Bonache, L., & Yokomori, T. (2004). Learning mild context-sensitiveness: Toward understanding children's language learning. *ICGI* (pp. 53–64).

de la Higuera, C. (2005). *Complexity and reduction issues in grammatical inference* (Technical Report ISSN 0946-3852). Universität Tübingen.