Negative Data in Learning Languages

Sanjay Jain* Efim Kinber†

Based on motivations from theories of language acquisition by children, Gold [6] developed an algorithmic model of learning (in the limit) from examples. This model may be described as follows. A learner receives as input, one by one, $x_0, x_1, \ldots$, where, $\{x_0, x_1, \ldots\}$ is exactly the target language, except possibly for a special pause symbol (which is useful for dealing with empty language). Note that there is no particular order among the elements $x_0, x_1, \ldots$, and repetitions are allowed. As the learner is receiving this data, it conjectures a sequence of grammars, $g_0, g_1, \ldots$ which are intended as descriptors of the target language. The learner can be regarded as successful, if eventually the sequence of grammar stabilizes to a grammar g which generates/enumerates/accepts/ the target language. This model of learning is called **TxtEx** in the literature (**Txt** stands for "text", which is a complete positive data presentation, and "**Ex**" stands for explanatory learning). Note that it is more interesting to consider learnability of a class of languages by a single learner (since, if we are only interested in learning one fixed language, then some learner — which just outputs the grammar for the fixed language — can easily learn it). The influence of Gold's paradigm [6] to human language learning is discussed by [15, 18, 17, 13].

Note that in the above model, the learner only receives elements of the language as input. It is not given any explicit information about elements not in the target language. This was based on the studies by linguists which hypothesised that children rarely, if ever, get negative information (see for example, [2, 7, 4]).

Along with the above model of learning from positive data, Gold also studied learning from both positive and negative data. In this model, a learner is given all elements of the language, one by one, marked as positive, as well as all non-elements of the language, one by one, marked as negative. This criteria of learning is called **InfEx**. However based on studies about child learning, it is unrealistic to expect that children get all the negative data. On the other hand, as some studies point out, [2, 7, 4], children do get something more than just positive data.

The aim of the current paper is to survey some models of learning, where small amounts of

*Supported in part by NUS grant number R252-000-127-112. School of Computing, National University of Singapore, Singapore 119260 sanjay@comp.nus.edu.sg.

†Department of Computer Science, Sacred Heart University, Fairfield, CT 06432-1000, U.S.A. Email: kinbere@sacredheart.edu

negative data is provided to the learner. We will first consider two models of small amounts of negative data provided to the learner. These models and results are based on work in [16, 5, 12, 1]. We will then consider the case where negative data is provided to the learner via counterexamples to their conjectures. This is based on the philosophy that parents often correct their children by providing them counterexamples. This part is based on work done in [8, 9]. We also introduce and briefly consider a model in which learners are provided with random negative examples.

Before we study different models of negative data, it is useful to also consider some variants of the basic model of learnability from text as described above. Case and Lynes [3] (see also [14]) studied the case where the final hypothesis of the learner may not accurate, but have upto n errors (finite number of errors). This criteria of learning is called $\mathbf{TxtEx}^n$ ($\mathbf{TxtEx}^*$). This was motivated by the fact that humans rarely, if ever, learn a language perfectly. Furthermore, Case and Lynes [3] (also see [14]) considered the case when learner need not syntactically converge to a grammar, but eventually output only correct grammars (i.e., sort of semantically converge rather than syntactically converge). In this model for all but finitely n, the grammar g_n is a grammar for the target language. This, criteria of learning is called **TxtBc**-learning. **Bc** here stands for behaviorally correct. $\mathbf{TxtBc}^n$ and $\mathbf{TxtBc}^*$ can be naturally defined.

Fulk [5] considered the case when, in addition to positive data, the learner is provided with a grammar for the complement of the language. Note that one can generate complete negative data using the grammar for the complement of the language. Fulk went on to show that this allows the learner to learn more than what can be learned using informants, that is using both complete positive and complete negative data. Though interesting, this model is quite unrealistic in the sense that children are definitely not given a grammar for the complement of the language. Most of the literature (see for example, [2, 7, 4]) also argues that children do not get complete negative data. What is more realistic is that a learner is provided with some negative data, probably carefully selected or based on what the child has learnt (that is in a way based on child's current conjecture). Jain and Sharma [10] considered a modification where the learner instead of being given a grammar for the complete $\overline{L}$, is given only for a subset of $\overline{L}$, where this subset satisfies some density constraints. Despite being somewhat weaker than Fulk's model, it still seems unrealistic to expect that children are provided with grammars for any parts of the complement of the target language.

Based on this, Shinohara [16] considered the case where the learner is given $\geq n$ (n fixed beforehand) arbitrary negative examples along with the complete positive data about the language. Clearly, this is possible only when complement of the language does contain at least n elements. Shinohara showed that this method of presenting negative data is not useful, in the sense that it does not give any learnability advantages over just positive data. Extending this work, Baliga, Case and Jain [1] considered the case that the learner is given

upto n carefully chosen elements of the complement of the language. These negative examples may be considered as *core* negative data. Intuitively, this was aimed to model the scenerio when a teacher carefully selects the negative examples to be provided to the student. Indeed, as expected this model turned out to be quite powerful. For example, it can be shown that the class of all recursively enumerable sets, $\mathcal{E}$, can be learned by some learner in **TxtEx** sense, when it additionally receives upto two carefully selected negative examples. Even one carefully selected negative example is enough if one allows upto one error in the final grammar, or allows behaviourally correct learning. In contrast one carefully selected negative example is not enough to learn the class $\mathcal{E}$ according to **TxtEx** crtieria, though it still can be shown to be quite useful.

The reason for this apparent gains by having only one or two negative examples in the above model is based on the fact that one can "code" information into these negative data, allowing the learner to essentially extract a grammar for the target language from the negative data. To avoid such coding, Baliga, Case and Jain (motivated by a model considered by Motoki [12]) considered the following modification. For each possible target language, besides the core negative data, the learner may be given some further negative data. This model of learning is called open negative data, reminding one of the basic open sets for the topology with respect to which enumeration operators are continuous. As the learner may not be able to distinguish core negative data from the other negative data, the effects of "coding" are somewhat eliminated. This model turned out to be quite useful in studying the effects of negative data. In particular, above criteria lie strictly between **TxtEx** and **InfEx** models of learning. Let $\mathbf{NegO}^n\mathbf{I}$ ($\mathbf{NegO}^*\mathbf{I}$) denote the criteria of learning formed when the core negative information is of size at most n (the core negative information is of finite size), and $\mathbf{I}$ is the basic model of learning (such as $\mathbf{Ex}^a$, or $\mathbf{Bc}^a$). It can be shown that $\mathbf{NegO}^*\mathbf{I}$ turns out to be of the same power as $\mathbf{InfI}$. Furthermore, each additional element allowed in the core, gives learnability advantages (that is $\mathbf{NegO}^{n+1}\mathbf{I}$ allows learning strictly more classes compared to $\mathbf{NegO}^n\mathbf{I}$). On the other hand, the finite negative core information is not enough to overcome extra errors (that is, one can learn something in $\mathbf{TxtEx}^{n+1}$ model of learning, but cannot in $\mathbf{NegO}^*\mathbf{TxtEx}^n$ model of learning). Additionally, it was shown that small packets of negative information also lead to increased *speed* of learning. This result agrees with a psycholinguistic hypothesis of McNeill correlating the availability of parental expansions with the speed of child language development. McNeill [11] posits that there is *faster* learning of language for children in homes in which more corrections (usually in the form of *possibly exemplary* expansions) are given. These corrections are, in part, a form of negative information.

Note that in both the model considered above, one selects carefully negative examples based on the language being learned. However, in reality often negative examples are formed more as "counterexamples" based on errors done by child, rather than being preselected. To

model such a scenerio, Jain and Kinber [8], considered a criteria of learning where the learner is given a negative counterexample to each of its conjectures, if it exists. This model of learning is called **NCEx**. This model turned out to be robust with respect to different variations (giving least counterexamples, or the counterexamples being delayed). Besides the usual hierarchy results showing the advantages of having counterexamples, the paper [8] contrasts this criteria with **TxtEx** and **InfEx**, showing that in some cases structurally it behaves more like "**InfEx**" rather than like "**TxtEx**". For example, results such as (a) if $\mathcal{L} \in \mathbf{NCEx}$ then so is $\mathcal{L} \cup \mathcal{S}$, for any finite class $\mathcal{S}$ of recursive languages, (b) $\mathbf{NCEx} \subseteq \mathbf{NCBc}$ follow more along the lines of results in learning from informants. On the other hand, it is shown that in some cases full negative data, informant, is needed for learning, and just counterexamples are not enough. A surprising result, in the case of behaviorally correct learning is that the whole class $\mathcal{E}$ can be learned in $\mathbf{NCBc}^1$ model — making it more powerful than even learning from informants! (by contrast $\mathbf{NCBc} \subset \mathbf{InfBc}$ and $\mathbf{NCEx}^a \subseteq \mathbf{InfEx}^a$).

An interesting complexity aspect is that, for **Ex** model, though **NCEx** is a strict subset of **InfEx**, it can sometimes give huge complexity advantages. That is, in some cases one can learn a class in **NCEx** model using only n mind changes, whereas learning with informants requires exponentially many mind changes. In a variation of **NCEx** model, where least negative counterexamples are given, one can even show that there are classes which are learnable using 1 mind change, though learning with informants requires unbounded number of mind changes! Though, as mentioned above, various variations of negative counterexample models do not give different learning power, there is often complexity advantages which may result from a particular variation.

Learning from counterexamples also addresses a general concern about overgeneralization in learning. When one only receives positive data, then overgeneralized hypothesis cannot be corrected based on input data alone. However, if negative counterexamples are provided to the learner, then one can address this issue.

One can view getting counterexamples, as asking a "subset query" about the conjecture to a teacher. However in the usual model of learning from subset queries, a learner is allowed to query about other languages (besides just the conjectured language) being subsets of target language. This led [9] to consider learning with subset (and other kind of) queries. It can be shown that if a **TxtEx** learner is allowed finitely many (but unbounded) subset queries, then the learning ability is same as that in the **NCEx** model. If the learner is allowed infinitely many subset queries, then a learner (using texts) can learn all the recursively enumerable languages. Thus it is more interesting to study the case when the number of queries is bounded. Jain and Kinber showed several results comparing the criteria of learning with negative counterexamples and subset queries, and giving hierarchies based on number of queries allowed. They also showed hierarchies based on variations of the query model where no answers are provided with least, arbitrary, or no counterexamples.

An interesting research work to consider would be to see how random negative examples work — this may be more closer to how humans learn languages. It can be shown that often random negative examples do help.

Список литературы

[1] G. Baliga, J. Case, and S. Jain. Language learning with some negative information. *Journal of Computer and System Sciences*, 51(5):273–285, 1995.

[2] R. Brown and C. Hanlon. Derivational complexity and the order of acquisition in child speech. In J. R. Hayes, editor, *Cognition and the Development of Language*. Wiley, 1970.

[3] J. Case and C. Lynes. Machine inductive inference and language identification. In M. Nielsen and E. M. Schmidt, editors, *Proc. of the 9th ICALP*, volume 140 of *LNCS*, pages 107–115. Springer-Verlag, 1982.

[4] M. Demetras, K. Post, and C. Snow. Feedback to first language learners: The role of repetitions and clarification questions. *Journal of Child Language*, 13:275–292, 1986.

[5] M. Fulk. *A Study of Inductive Inference Machines*. PhD thesis, SUNY/Buffalo, 1985.

[6] E. M. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[7] K. Hirsh-Pasek, R. Treiman, and M. Schneiderman. Brown and Hanlon revisited: Mothers' sensitivity to ungrammatical forms. *Journal of Child Language*, 11:81–88, 1984.

[8] S. Jain and E. Kinber. Learning language from positive data and negative counterexamples. In John Case Shai Ben-David and Akira Maruoka, editors, *ALT' 2004*, volume 3244 of *LNAI*, pages 54–68. Springer-Verlag, 2004.

[9] S. Jain and E. Kinber. Learning languages from positive data and a finite number of queries. In Kamal Lodaya and Meena Mahajan, editors, *FST&TCS'04*, volume 3328 of *LNCS*, pages 360–372. Springer-Verlag, 2004.

[10] S. Jain and A. Sharma. Learning in the presence of partial explanations. *Information and Computation*, 95:162–191, 1991.

[11] D. McNeill. Developmental psycholinguistics. In F. Smith and G. Miller, editors, *The Genesis of Language*, pages 15–84. MIT Press, 1966.

[12] T. Motoki. Inductive inference from all positive and some negative data. *Information Processing Letters*, 39(4):177–182, 1991.

[13] D. Osherson, M. Stob, and S. Weinstein. *Systems that Learn: An Introduction to Learning Theory for Cognitive and Computer Scientists*. MIT Press, 1986.

[14] D. Osherson and S. Weinstein. Criteria of language learning. *Information and Control*, 52:123–138, 1982.

[15] S. Pinker. Formal models of language learning. *Cognition*, 7:217–283, 1979.

[16] T. Shinohara. *Studies on Inductive Inference from Positive Data*. PhD thesis, Kyushu University, Kyushu, Japan, 1986.

[17] K. Wexler. On extensional learnability. *Cognition*, 11:89–95, 1982.

[18] K. Wexler and P. Culicover. *Formal Principles of Language Acquisition*. MIT Press, 1980.