Getting Order Independence in Incremental Learning

Antoine CORNUEJOLS

Equipe Inférence et Apprentissage
Laboratoire de Recherche en Informatique (LRI), UA 410 du CNRS
Université de Paris-sud, Orsay
Bâtiment 490, 91405 ORSAY (France)
email (UUCP) : antoine@lri.lri.fr

Abstract[1]

Except for very special conditions, incremental learning systems are order dependent. Ideally, a well informed teacher could thus use this property to guide the learner towards some target concept and away from others. This would be equivalent to constraining learning using prior knowledge about the learner and on the task at hand.
In order to better define what form this prior knowledge could take, this paper studies the case where incremental learning is order independent.

1. Introduction

Ordering effects in Incremental Learning have been widely mentioned in the literature without, however, being the subject of much specific study except for some rare pioneering works [1,4,8]. In short, ordering effects are observed when, given a collection of data (e.g. examples in inductive concept learning), different ordered sequences of these data lead to different learning results. Ordering of data therefore seems to be equivalent to a preference bias acting among all the models or hypotheses that the learning systems could have reached given the collection of data (if they had been presented in every possible orders). Hence, it undoubtedly amounts to some additional knowledge supplied to the system.

Learning without some bias that allows the reduction of the search space for the target concept or model is impossible except in the crudest form of rote learning. When looking more closely, it is usual to distinguish between :

- *representation bias* : where the search space is constrained because all partitions of the example space can not be expressed in the hypothesis space considered by the system (this is the basis for inductive generalization and is the main topic of current Machine Learning theory [10]), and
- *preference bias* : which dictates which subspace should be preferred in the search space (e.g. prefer simple hypotheses over more complex ones) (this type of bias has been much less studied because it touches on procedural aspects instead of declarative ones only).

Because ordering of inputs allows one to favor some models over some others, it seems to amount to a preference bias that chooses between competing hypotheses. In spite of this resemblance however, there is a deep difference with the biases generally discussed in Machine Learning. Indeed, with ordering effects, one *observes* the preference but cannot pinpoint directly where in the learning system it lies and how it works. This is in contrast with what is considered classically as a bias, where one can identify *operational* constraints _e.g. isolate representation constraints or procedures for choice between hypotheses. Thus we use the term global preference bias to denote preference among models due to ordering effects *after* a sequence of inputs has been observed, and the term local preference bias to denote the local choice strategy followed by the system when at each learning step it must choose to follow some paths and discard others.

Two questions then immediately come up :

1- *What is the relationship between a global preference bias and a local one ?*

2- *What is the relationship between a global preference bias that is observed or aimed at and a corresponding teaching strategy that specifies the order of inputs ? In other words, how to design a teaching strategy so as to get a certain global preference bias ?*

In order to answer these questions, and particularly the first one, it is necessary to determine the **causes of the ordering effects**.

[1] A fuller version of this paper with complete proofs is available as [2]

In this regard, it is instructive to look at incremental learners that are NOT order dependent, like the candidate elimination (CE) algorithm in Version Space[9], ID5 [14], or systems that are not usually considered as learning systems but could be, such as TMS [3] or some versions of the Bayesian Inference nets of Pearl [11]. They all have in common that they do not forget any information present in the input data. Thus, even when they make a choice between alternative hypotheses, like ID5 or TMS and unlike the CE algorithm, they keep enough information to be able to compare all potential competing models so as to select the best one at any moment, and change their mind if needed. They are therefore equivalent to non-incremental learning systems that get all the data at once and focus on the best hypothesis given the information supplied.

Forgetting of information lies therefore at the heart of order dependence in incremental learning. But forgetting can take two faces. In the first one, information present in the input data is lost, meaning that the current hypothesis space considered by the learner is underconstrained. In the second one, by contrast, what is lost are potential alternatives to the current preferred hypotheses, which amounts to overconstraining the space of possibilities. **This last form of forgetting is equivalent to a local preference bias** which chooses among competing hypotheses which ones to pursue.

This raises then a more specific question than the aforementioned ones, but which contributes to the same overall goal :

3. *In which case an incremental learner can be order independent ? Or, in other words, which information can be safely forgotten without altering the result of learning whatever is the ordering of inputs ?*

It is this last question that this paper focuses on. It must be kept in mind that it is equivalent to the question : what local preference bias leads to a null global preference bias (i.e. to order independence) ?

In the following of the paper, **we will restrict ourselves to a simple concept learning model** in which the learner attempts to infer an unknown target concept f, chosen from a known concept class F of $\{0,1\}$-valued functions over an instance space X. This framework allows us, in the next section, to define a measure of the information gained by the learning system and of the effect of a local bias on this information. This measure naturally suggests an equivalence relation between local preference bias and additional instances, which is detailed in section 3. Then, in section 4, it becomes a relatively simple matter to answer question 3 above. The conclusion compares the framework adopted here with the emerging one of teachability and discusses the results obtained.

2. Information measure and local preference bias

In this section, we are interested in formalizing and quantifying the effect of a local preference bias on what is learned by the system. For this, we first define a characterization of the information maintained by a learner.

Let F be a concept class over the instance space X, and $f \in \mathsf{F}$ be a target concept. The teacher has a collection of examples $EX=\{x_i f(x_i)\}$ at his disposal, and makes a sequence $\mathbf{x}=x_1,x_2,\ldots,x_m,x_{m+1},\ldots$ with $x_m \in EX$ for all m. The learner receives information about f incrementally via the label sequence $f(x_1),\ldots f(x_m)f(x_{m+1}),\ldots$ For any $m \geq 1$, we define (with respect to $\mathbf{x},f$) the mth *version space* :

$$\mathsf{F}_m(\mathbf{x},f) = \{\hat{f} \in \mathsf{F} : \hat{f}(x_1) = f(x_1),\ldots,\hat{f}(x_m) = f(x_m)\}$$

The version space at time m is simply the class of all concepts in F consistent with the first m labels of f (with respect to $\mathbf{x}$). $\mathsf{F}_m(\mathbf{x},f)$ will serve as a **characterization of what is known to the learner at time m about the target concept f.**

We know from Mitchell [9] that the version space can be economically represented and stored using the boundary sets S-set (set of the most general hypotheses that are more specific than the concepts in the version space), and G-set (set of the most specific hypotheses that are more general than the concepts in the version space). Each new example $(x_m f(x_m))$ provides new information if it allows to reduce the version space by modifying, through the CE algorithm, either one of the boundary sets. Generally, the S-set and the G-set contains many elements, and in worst cases, they can grow exponentially over some sequences of examples [6].

A local preference bias is a choice strategy which, at any time m, discards parts of the current version space, generally in order to keep the boundary sets manageable. In this way, it reduces the version space and acts as if there had been some additional information that had allowed to constrain the space of hypotheses. The next section gives a closer look at this equivalence.

3. Bias and additional instances

We assume that the incremental learner maintains a version space of potential concepts by keeping the boundary sets. We assume further that the local preference bias, if any, acts by removing elements of the S-set and/or of the G-set, thus reducing the version space. Indeed, in so doing, it removes from the version space all concepts or

hypotheses that are no longer more general than some element of the S-set and more specific than some element of the G-set. Besides, the resulting version space keeps its consistency since, in this operation, no element of the resulting S-set is more general than other elements of the S-set or of the G-set, and vice-versa, no element of the G-set can become more specific than other elements of the G-set or of the S-set.

Theorem 1 : *With each choice it makes, the local preference bias acts as if additional examples had been known to the learner.*

Proof : (i) Case of the reduction of the S-set. For each element g_i of the S-set it is possible to find additional examples which, if considered by the CE algorithm, would lead to its elimination of the S-set. It suffices to take positive instances that are more specific than the other elements of the S-set but not than g_i. In this way, the CE algorithm would generalize g_i just enough to cover the new instances, and this would result in an element of the S-set that would be more general than the others, hence eliminated.

(ii) Case of the reduction of the G-set. In the same way, in order to eliminate an element g_i of the G-set through the CE algorithm, it suffices to provide negative instances covered by g_i but not covered by the other elements of the G-set and by the S-set. The CE algorithm would then specialize just enough to exclude the negative instances, and this would result in an element of the G-set that would be more specific than others, hence eliminated. ❑

The key insight that makes this theorem interesting is that, thanks to it, we can transform a problem involving the comparison of the result of the local preference bias acting along different sequences of examples $\mathbf{x_j}$ into a problem where there is no local preference bias, thus no different histories, but a set of additional examples that are provided to the CE algorithm. This allows us to obtain the results of section 4.

4. Bias and order independence

In this section, we will study what kind of local preference bias a learner can implement so as to stay order independent. We assume that the teacher has a set of n examples EX, and draws a sequence $\mathbf{x}$ of these according to her requirements.

Furthermore, we assume order independence, i.e. :

(1) $\forall \mathbf{x}, F_n^{LB}(\mathbf{x},f) = C$, where C is constant.

(We use the notation F_n^{LB} to differentiate a learner implementing a local bias (LB) from one that does not and only implements the CE algorithm).

Theorem 2 : *An incremental learner implementing a local preference bias is order independent for a collection EX of instances if the action of this bias is equivalent for all possible sequences $\mathbf{x}$ of elements of EX to the supply of the same set of additional instances.*

Proof : It follows immediately from theorem 1. ❑

Remark : This is a sufficient condition, theorem 3 gives necessary conditions.

Let S and G be respectively the S-set and the G-set that the CE algorithm would obtain from the collection of instances EX, and let S_C and G_C be respectively the S-set and the G-set of C in (1) (the version space obtained on any sequence $\mathbf{x}$ of EX by the learner implementing the local preference bias).

Theorem 3 : *For an incremental learner implementing a local preference bias to be order independent for EX leading to the version space C, it is necessary that the action of this bias be equivalent to the supply of :*

- a set of positive instances such that each one is covered by all elements of S_C and by all but one of the elements of S not covered by S_C, and for each elements of S not covered by S_C there is an instance covered by all elements of S and by all others elements of S not covered by S_C, and ,

- a set of negative instances such that each one is covered by one element of G not covering any element of G_C and not covered by any other element of G and of G_C, and for each element of G not covering G_C there is an instance covered by this element and not by any other element of G or of G_C.

Proof : It follows easily from theorem 1. See [2]. ❑

The next theorem is the application of theorem 3 to the case where $C = F_n(EX,f)$, that is the case where the local preference bias leads to the null global preference bias, i.e. has no effect. Such a local bias can be seen as eliminating options judiciously since the result obtained after any sequence $\mathbf{x}$ of EX is the same as what the CE algorithm would get on EX. In this case S and S_C are one and the same as are G and G_C.

Theorem 4 : *For an incremental learner implementing a local preference bias leading to the same result as an incremental learner without a local bias, it is necessary that the action of this bias be equivalent to the supply of :*

- a set of positive instances such that each one is covered by all elements of S_C, and ,

- a set of negative instances such that each one is covering all elements of G_C.

It is as if this local preference bias knew "in advance" the collection EX of instances, and eliminated elements of the

S-set and of the G-set judiciously. This leads to the final theorem.

Theorem 5 : *It is not possible for a local deterministic preference bias to lead to a null global preference bias for any arbitrary collection EX of examples.*

Proof : Indeed, this would mean that the learner, through its preference bias, was always perfectly informed in advance on the collection EX of examples held by the teacher. ❑

5. Conclusion

In this research, we are interested in the following **general question** : *given a collection of examples (or data in general), how can a teacher, a priori, best put them in sequence so that the learner, a deterministic incremental learning system that does not ask questions during learning, can acquire some target concept (or knowledge)?*

This question, that corresponds to situations where the teacher does not have the choice of the examples and can not interpret the progress made by the student until the end of the learning period, leads to the study of incremental learning per se, independently of any particular system.

This framework is **to be compared with** the recent surge of interest for "teaching strategies" that allow to optimally teach a concept to a learner, thus providing lower bounds on learnability complexity [5,13]. The difference with the former framework is that in one case the teacher can only play on the order of the sequence of inputs, whether in the other case, the teacher chooses the most informative ideal examples but does not look for the best order (there are some exceptions such as [12]).

This paper has outlined some **first results** concerning order sensitivity in supervised conceptual incremental learning. The most important ones are : (i) that order dependence is due to forgetting of possibilities corresponding to a local preference bias that heuristically selects the most promising hypotheses, (ii) that this bias can be seen as the result of additional instances given to the learner (i.e. prior knowledge built into the system), (iii) that (ii) allows to replace the difficult problem of determining the action of the local bias along different sequences of instances by a problem of addition (which is commutative, i.e. order independent) of instances to an order independent learner, which leads to (iv) that there are strong contingencies for an incremental learner to be order independent on some collections of instances (either the corresponding prior knowledge is well-tailored to the future potential collections of inputs, or there is no prior knowledge, thus no reduction of storage and computational complexity).

Questions for future research include : are these results extensible to more general learning situations (e.g. unsupervised) ? given a local preference bias, how to determine a good sequence ordering so as to best guide the system towards the target knowledge ?

Acknowledgments : I thank all the members of the Equipe Inference et Apprentisssage, and particularly Yves Kodratoff, for the good humoured and research conducive atmosphere so beneficial to intellectual work.

References :

[1] **CornuéJols** A. (1989) : *"An Exploration into Incremental Learning : the INFLUENCE System"*, in Proc.of the 6th Intl. Conf. on Machine Learning, Ithaca, June 29-July 1, 1989, pp.383-386.

[2] **CornuéJols** A. (1992) : *"Getting Order Independence in Incremental Learning"*. (In preparation)

[3] **Doyle** J. (1979) : *"A truth maintenance system"*. Artificial Intelligence, 12, 231-272, 1979.

[4] **Fisher, Xu & Zard** (1992) : *"Ordering Effects in COBWEB and an Order-Independent Method"*. To appear in Proc. of the 9th Int. Conf. on Machine Learning, Aberdeen, June 29-July 1st, 1992.

[5] **Goldman & Kearns** (1991) : *"On the Complexity of Teaching"*. Proc. of COLT'91, Santa Cruz, Aug. 5-7 1991, pp.303-314.

[6] **Haussler** D. (1988) : *"Quantifying inductive bias: AI learning algorithms and Valiant's learning framework"*. Artificial Intelligence, 36, 177-222.

[7] **Hirsh** H. (1990) : *"Incremental Version-Space Merging"*. Proc. of the 7th Int. Conf. on Machine Learning, Univ. of Austin, Texas, June 21-23, 1990, pp.330-338.

[8] **MacGregor** J. (1988) : *"The Effects of Order on Learning Classifications by Example: Heuristics for Finding the Optimal Order"*, Artificial Intelligence, vol.34, pp.361-370, 1988.

[9] **Mitchell** T. (1982) : *"Generalization as Search"*, Artificial Intelligence, vol.18, pp.203-226, 1982.

[10] **Natarajan** B. (1991) : *Machine Learning. A Theoretical Approach*. Morgan Kaufmann, 1991.

[11] **Pearl** J. (1988) : *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[12] **Porat & Feldman** (1991) : *"Learning Automata from Ordered Examples"*. Machine Learning,7, pp.109-138, 1991.

[13] **Salzberg, Delcher, Heath & Kasif** (1991) : *"Learning with a helpful teacher"*. Proc. of the IJCAI-91, pp.705-711.

[14] **Utgoff** P. (1988) : *"ID5 : An Incremental ID3"*. In Proc. of the 5th Intl. Conf. on Machine Learning, Ann Arbor, June 12-14, 1988, pp.107-120.