Computational Science and Data Mining

Flaviu Adrian Mărginean

Department of Computer Science, The University of York
Heslington, York YO10 5DD, United Kingdom
flav@cs.york.ac.uk

Abstract. The last decade has witnessed an impressive growth of Data Mining through algorithms and applications. Despite the advances, a computational theory of Data Mining is still largely outstanding. This paper discusses some aspects relevant to computation in Data Mining from the point of view of the Machine Learning theoretician. Computational techniques used in other fields that deal with learning from data, such as Statistics and Machine Learning, are potentially very relevant. However, the specifics of Data Mining are such that most often those techniques are not directly applicable but require to be re-cast and re-analysed within Data Mining starting from first principles. We illustrate this with a PAC-learnability analysis for a Data Mining-like task. We show that accounting for Data Mining specific requirements, such as inference of weak predictors and agnosticity assumptions, requires the generalisation of the classical PAC framework in novel ways.

1 Introduction

Data Mining is a relatively recent research field, formed over the last ten years at the intersection of Database Theory, Statistics and Machine Learning [12]. It is perhaps surprising, given the three parent fields, that the theoretical foundations of Data Mining are still in the incipient development stages. Databases have been studied since the '60s, while Statistics is century-old. Machine Learning has existed as a problem since Turing's seminal work in the early '50s. All three fields have by now acquired solid theoretical foundations and can be regarded as mature fields. By way of contrast, not enough theoretical understanding exists about the nature and purposes of Data Mining, and the mere fact that ONE mythical theory of Data Mining is still being searched for is, perhaps, a sign of the immaturity of the field. The efforts to develop full-fledged foundations for Data Mining might have been hampered by the so-called reductionist approaches [14], i.e. those aiming to reduce Data Mining to one or another of its three parents mentioned above. The difficulty with these approaches seems to lie with the fact that, while some aspects of Data Mining fit neatly in either Database Theory, Statistics or Machine Learning, neither of the parents alone can provide a fully satisfactory account of all the more important characteristics of Data Mining. More problematically, even when the general methodology can be traced to one

P.M.A. Sloot et al. (Eds.): ICCS 2003, LNCS 2659, pp. 644–651, 2003.

of the parent fields, it may be the case that the relevant techniques or tools are not directly applicable, but require to be re-invented instead, within Data Mining, starting from first principles.

One such problem is the issue of learnability in Data Mining. An early preoccupation of Machine Learning theoreticians has been to assign meaning to the learning process, i.e. to build computational models of learning. A mathematical model of learning should be able to tell us something about the learning target, as well as about the computational resources needed in order to learn successfully (information, time, space). In contrast, in Data Mining one usually searches for the unknown and unexpected: *"unknown and unexpected patterns of information" (Parsaye), "previously unknown relationships and patterns within data" (Ferruzza), "valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad), "previously unknown, comprehensible, and actionable information" (Zekulin), "unsuspected relationships in observational data sets" (Hand-Mannila-Smyth, [6]).* A direct application of computational models developed for Machine Learning is therefore not possible, for at least two reasons. First, a Data Mining task may consist of finding many weak predictors in a hypothesis space whereas in Machine Learning one strong predictor is normally sought [4]. This is the so-called *local-global* problem [13]. Secondly, Machine Learning algorithms usually make an assumption of representability of the data-generating mechanism within the hypothesis space of the learner. This is certainly not the case within Data Mining, wherein the lack of assumptions regarding the phenomenon generating the data is a matter of principle as described above. This is the so-called *agnosticity* problem [17].

The purpose of this paper is to give a computational treatment of learning in Data Mining and a PAC-learnability analysis. Such analyses have not so far been done, to our knowledge. We wish to emphasise that this is not the only possible computational treatment of learning in Data Mining. Many computational models of learning exist in the theoretical literature on Machine Learning (PAC-learning, U-learning, etc.) as well as many forms of learning (inductive, analytical, Bayesian, reinforcement, etc.). There is no a priori reason to expect the situation to be any different in Data Mining. Rather we wish to showcase a type of computational analysis, which takes account of the specifics of Data Mining as mentioned above. The paper is organised as follows. In Section 2 we provide some technical preliminaries needed in the rest of the work: a description of Valiant's PAC-learnability framework in Section 2.1; a brief description of Mitchell's agnostic learning in Section 2.2; and a presentation of the local-global problem in Section 2.3. Section 3 is the main section of the paper, containing the definitions necessary for our extended PAC framework and the main result. Finally, Section 4 concludes with a discussion and gives some directions for further work.

2 Background

2.1 PAC-Learnability Framework

We first present the PAC model of learning, generally following Mitchell [17].

In the PAC model of learning (whereof many variants exist), we are customarily given an instance space X and two subsets of the powerset of X, the concept space $\mathcal{C}$ and the hypothesis space $\mathcal{H}$: $\mathcal{C}, \mathcal{H} \subseteq 2^X$; they can equally well be thought of as spaces of Boolean-valued functions over X. These two spaces are usually implicitly defined by some representational schemes, for instance DNF or CNF formulae, first-order representations etc. However, this will not be important in our analysis. It is assumed in the classical model that $\mathcal{C} \subseteq \mathcal{H}$, i.e. any possible target is representable in the hypothesis space of the learner. This entails the possibility of finding a unique strong predictor in the hypothesis space.

In Data Mining, we no longer make this rather strong assumption. Models of learning that do not make this assumption have been described before under the name agnostic [17] or robust learning [8], however those models differ in a number of ways from the treatment that we propose here.

It is further assumed in the PAC model that a distribution $\mathcal{D}$ over X is given, unknown to the learner but fixed in advance. The purpose will be to probably approximately learn $c \in \mathcal{C}$ by querying an oracle that makes random independent draws from the distribution $\mathcal{D}$. Every time the oracle is queried by the learner, it draws an instance $x \in X$ at random according to $\mathcal{D}$ and returns the pair $\langle x, c(x) \rangle$ to the learner. An approximation $h \in H$ to $c \in \mathcal{C}$ is evaluated with respect to the distribution $\mathcal{D}$ over X: the error of the approximation is the probability that an instance from X randomly drawn according to $\mathcal{D}$ will be misclassified by h. It is required that a learner, using reasonable amounts of computational and informational resources (time, space, queries to the random oracle), output a hypothesis that with high confidence approximates the target well-enough. The use of resources such as time and space define the computational complexity of PAC learning, while the number of queries to the random oracle needed to probably approximately infer the unknown target defines the sample complexity or information-theoretic complexity of the PAC learning. It is the latter we will be concerned with in this paper.

2.2 Agnostic Learning

In agnostic learning [17, 8], no prior assumption is made to the effect that $\mathcal{C} \subseteq \mathcal{H}$. Since the process generating the data is not assumed to be representable in the hypothesis space of the learner, in agnostic learning one seeks the hypothesis with the smallest error over the training data, h^D_{best}. It is then shown using Hoeffding bounds that the true error of this hypothesis will not exceed an ϵ overhead compared to the error over the training data. This framework is not fully satisfactory and cannot be applied directly to the study of Data Mining. It does not appear clear in agnostic learning why the hypothesis h^D_{best} would be interesting for us from the point of view of learning the target c. Whilst it is guaranteed that its true error will be not much bigger than its error over the

training data, it is not clear why this would be enough justification for singling out h^D_{best}. Its error over the training data might be smallest of all hypotheses, but it can still be too big for h^D_{best} to be a reasonable predictor for c. If we however accept that we are also interested in weak predictors, and not necessarily in strong predictors, it is not clear why we should pick just one of them. Rather, from Data Mining we know that we are usually interested in any number of weak predictors satisfying some quality criteria. This is the object of Section 2.3.

2.3 Local and Global Issues in Data Mining

Data Mining may be concerned [13] with one of two problems. The probabilistic modelling research tradition views Data Mining as the task of approximating a global model underlying the data in the form of a joint distribution.

The other approach can be seen as as *"the essence of Data Mining — an attempt to locate nuggets of value amongst the dross" ([5, Hand])*. The typical example is the discovery of frequently occurring patterns, wherein patterns and their associated frequencies give local properties of the data that can be understood without having information about the global mass of the data. The collection of such patterns may be used as a first step towards the global analysis of the data, whereby the collection of patterns collected in the mining phase undergoes various aggregation operations aimed at building a global description of the data. However, global approaches more genuinely belong to Machine Learning, where the assumption of a global generative process behind the data is better supported. Most prominent example of work in the local tradition is the research on association rule mining [13, 4, 1, 15]. We shall be concerned in this paper with the local problem. The trouble with building weak predictors (models) of the data, based on local information, is that of sampling. How can we make inductive leaps from training data to weak predictors when *"selecting only a sample may discard just those few cases one had hoped to detect" ([5, Hand])*? This problem, to our knowledge, has not been given a theoretical treatment so far.

3 PAC-Learnability Analysis for Mining

In this section we present our PAC framework for mining analysis, paying attention to the necessity to model both requirements specific to the mining process: *agnosticity* as described in Section 2.2 and *locality* as described in Section 2.3. Given that we no longer can rely on the assumption $\mathcal{C} \subseteq \mathcal{H}$ which would allow us to define the version space of the hypotheses consistent with the data, we have to find another way of defining interesting hypotheses from the point of view of consistency with the data. We do so by introducing the following quasi-order on $\mathcal{H}$, which is relative to the training data D and target c : $h_1 \succeq_{\mathcal{OD}(c,D)} h_2$ *iff* $\forall \langle x, c(x)\rangle \in D : \; h_2(x) = c(x) \Rightarrow h_1(x) = c(x)$. The order is subscripted $\mathcal{OD}$, indicating *desirability* of hypotheses in $\mathcal{H}$ (or Order of Desire).

It is trivial to verify that the quasi-order axioms (transitivity, reflexivity) are satisfied. The notion of Version Space $VS_{\mathcal{H},D}$ [17] generalises as follows:

$$VS_{\mathcal{H},D} \stackrel{\text{def}}{=} \{h \in H \mid h|_D = c|_D\}$$

$$SVS_{\mathcal{H},D} \stackrel{\text{def}}{=} \{h \in H \mid \nexists\, h' \in \mathcal{H} \textit{ such that } h' \succ_{\mathcal{OD}(c,D)} h\}$$

We call the generalised Version Space, Soft Version Space $SVS_{\mathcal{H},D}$. The word *soft* we use to indicate the graceful degradation of hypotheses with respect to consistency. With classical Version Spaces, hypotheses are classified in a crisp manner: hypotheses are either in (consistent) or out (inconsistent) of the Version Space. Soft Version Spaces $SVS_{\mathcal{H},D}$ retain the hypotheses maximally consistent with the data D, but the gap between them and the hypotheses left out of the version space is not as dramatic as in the classical case. Rather, consistency comes in degrees and there may be all sorts of shades, i.e. hypotheses that are more or less consistent with the data D according to the order $\mathcal{OD}(c, D)$. When $D = X$ we denote the Soft Version Space by $SVS_{\mathcal{H},c}$, the set of hypotheses in $\mathcal{H}$ maximally consistent with c over the entire instance space X. Let $\sim_{\mathcal{OD}(c,D)}$ be the equivalence relationship on $\mathcal{H}$ canonically induced by the quasi-order $\mathcal{OD}(c, D)$:

$$h_1 \sim_{\mathcal{OD}(c,D)} h_2 \quad \textit{iff} \quad h_1 \succeq_{\mathcal{OD}(c,D)} h_2 \textit{ and } h_2 \succeq_{\mathcal{OD}(c,D)} h_1$$

For the case $c \in \mathcal{H}$, the partial order $\mathcal{OD}(c, D)/\sim_{\mathcal{OD}(c,D)}$ on $\mathcal{H}/\sim_{\mathcal{OD}(c,D)}$ becomes the boolean order $B_1 \stackrel{\text{def}}{=} (0 < 1)$, with the subclass of consistent hypotheses corresponding to 1 and the subclass of inconsistent hypotheses corresponding to 0. Therefore, in this case the Soft Version Space becomes the set of consistent hypotheses, i.e. it reduces to the classical black-and-white definition of the Version Space $VS_{\mathcal{H},D}$. The following theorem establishes the sample complexity of "soft learning" for maximally consistent learners[1].

Theorem 1 (Soft Version Spaces are $\epsilon, \delta-$Exhaustible). *Let $\mathcal{C}, \mathcal{H} \subseteq 2^X$ be a concept space and a hypothesis space respectively, and let $VC(\mathcal{H}) < \infty$ be the finite Vapnik-Chervonenkis dimension of $\mathcal{H}$. For all $0 < \epsilon, \delta < 1$, for all $D \subseteq (X \times \{0, 1\})^m$ training data such that $m \geq \frac{1}{2\epsilon^2}(4\log_2(2/\delta) + 8\,VC(\mathcal{H})\log_2(13/\epsilon))$ and for all $h \in SVS_{\mathcal{H},c}$ there is $h' \in SVS_{\mathcal{H},c}$ and $h'' \in SVS_{\mathcal{H},D}$ such that, with probability at least $1 - \delta$, $error(h', c) \leq error(h, c) + \epsilon$ and $error(h'', h') < \epsilon$.*

Proof Outline. For reasons of space we only indicate the main steps of the proof's theorem. The first step is to show that for all $h \in SVS_{\mathcal{H},c}$ there is $h' \in SVS_{\mathcal{H},c}$ such that, with probability at least $1 - \delta$, $error(h', c) \leq error(h, c) + \epsilon$. This is done by restricting h to D and choosing h' such that $h' \succeq_{\mathcal{OD}(c,D)} h$ (one such maximal element with respect to $\mathcal{OD}(c, D)$ is bound to exist). It is also possible to show that h' can be so chosen that not only does $h' \in SVS_{\mathcal{H},D}$ but also $h' \in SVS_{\mathcal{H},c}$. However such a choice will necessarily be non-constructive. h'

[1] Compare with similar results in [17] regarding the complexity of PAC learning for consistent learners.

will do at least as well as h on the training data; therefore, by using Hoeffding bounds, we can bound with high confidence the true error of h' versus the true error of h. The second step involves showing that h' is probably approximable by an $h'' \in SVS_{\mathcal{H},D}$ with high confidence. This is essentially done by taking h' as a target, showing that there are elements in $SVS_{\mathcal{H},D}$ that are consistent with h' on the training data D, and applying the classical result of PAC-learnability from [17]. Unlike in the case of h', such elements can be chosen effectively, in effect as any $h'' \succeq_{\mathcal{OD}(c,D)} h$. Fewer examples would be needed for the second step, only $\frac{1}{\epsilon}(4\log_2(2/\delta) + 8\,VC(\mathcal{H})\log_2(13/\epsilon))$ according to the classical result for Version Spaces. However, this would only guarantee that elements in $SVS_{\mathcal{H},D}$ probably approximate elements in $SVS_{\mathcal{H},c}$, rather than ensuring that all elements in $SVS_{\mathcal{H},c}$ are probably approximately learned in the reasonable sense described by the theorem's statement. ■

4 Discussion and Further Work

A learnability analysis for Data Mining has been presented within the general methodology of Valiant's PAC-learning framework. The analysis shows that this type of analysis is feasible and that the process of mining is meaningful: the weak predictors $SVS_{\mathcal{H},D}$ inferred by maximal consistency from a polynomial sample have predictive power in a well-defined way. Moreover, the collection of all these predictors approximates the true target $SVS_{\mathcal{H},c}$, thereby collectively giving some global information about the data — just as practitioners of Data Mining would expect [13]. As far as the author is aware, this is the first learnability analysis for a Data Mining type of task.

There are various ways in which this work can be extended. First, we have investigated learnability with respect to only one resource, i.e. sample size. A mining algorithm will need to behave well not only in respect of the informational resources it requires (sample complexity), but also from the point of view of the hardware and time requirements (computational complexity). In other words, we have proved in this paper that any maximally consistent learner is effective, but there remains to be proved that there are such learners for given hypotheses spaces that are also efficient. Secondly, other models of learning may also be considered, as the PAC learning, although most common, is far from being the only learning model. It is, however, the simplest and best understood and it is the model of choice for first-time analyses of an inductive learning process [2, 10, 7].

We see the importance of this paper in conveying a principle: computational analyses of Data Mining problems and algorithms are possible, provided one re-develops the relevant techniques in the specific context of Data Mining rather than attempting a blind translation of techniques from other fields that deal with learning from data. Furthermore, the computational setting in this paper suggests new algorithms based on maximal consistency computation.

Traditionally, the APRIORI algorithm and its variants [1, 15] have handled the boolean inductive query evaluation problem with respect to single monotonic

constraints (e.g. minimum frequency). More recently, extensions have been proposed [3] that combine APRIORI with data structures based on VERSION SPACES in order to evaluate boolean inductive queries defined as conjunctions of both monotonic and anti-monotonic constraints. Those formalisms work well when the boolean inductive query evaluates to a non-empty set. However, natural computational settings exist wherein this is not a reasonable expectation for most of the queries. It is clear, in such settings, that the database still possesses some structure and some answers are better than others; for instance, one answer may satisfy more constraints in the inductive query than another. In cases where the classical formalism returns an empty answer set, we would instead be interested in computing those answers that most closely satisfy the inductive query. This problem requires the extension of the classical APRIORI formalism at both the conceptual and algorithmic level. In this paper we defined a framework entitled SOFT VERSION SPACES that can be viewed as describing the optimal 'soft match' $SVS_{\mathcal{H},D}$ between a language of patterns $\mathcal{H}$, herein more generally regarded as a set of hypotheses, and an inductive query D consisting of a conjunction of monotonic and anti-monotonic constraints over $\mathcal{H}$, herein more restrictedly viewed as the conjunction of the *h-membership* relations for positive and negative c-examples, respectively. This can be shown to generalise the classical APRIORI-based formalism in a natural way. The development of this idea and of a suitable SOFT-APRIORI algorithm for computing $SVS_{\mathcal{H},D}$ are topics for a future paper.

Acknowledgements.
The author thanks the anonymous reviewers for their comments. My mother, Maria Adriana (Puşi), and brother, Emil Raul, have provided moral support and encouragement during the writing of this paper. In loving memory of my departed father, Emil.

References

1. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo. Fast Discovery of Association Rules. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advance in Knowledge Discovery and Data Mining*, pages 307–328. AAAI Press/MIT Press, 1996.
2. M. Anthony and N. Biggs. *Computational Learning Theory*. Cambridge University Press, 1997.
3. L. De Raedt, M. Jaeger, S. Lee, and H. Mannila. A Theory of Inductive Query Answering. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)*, Maebashi, Japan, December 9–12 2002. Extended abstract.
4. D. Gunopulos, H. Mannila, R. Khardon, and H. Toivonen. Data Mining, Hypergraph Transversals, and Machine Learning (extended abstract). In *Proceedings of the* 16^{th} *ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 209–216, Tucson, Arizona, USA, 1997. ACM Press.

5. D.J. Hand. Statistics and Data Mining: Intersecting Disciplines. *ACM SIGKDD Explorations*, 1(1):16–19, 1999.
6. D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, 2001.
7. D. Haussler. Quantifying inductive bias: AI learning algorithms and Valiant's learning framework. *Artificial intelligence*, 36:177 – 221, 1988.
8. D. Haussler, S. Ben-David, N. Cesa-Bianchi, and P. Long. Characterizations of Learnability for Classes of {0,...,n}-valued Functions. *J. Comp. Sys. Sci.*, 50(1):74–86, 1995.
9. H. Hirsh. *Incremental Version Space Merging: A General Framework for Concept Learning*. Kluwer, 1990.
10. Michael J. Kearns and Umesh V. Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, 1994.
11. N. Lavrač, D. Gamberger, and V. Jovanoski. A Study of Relevance for Learning in Deductive Databases. *Journal of Logic Programming*, 40(2/3):215–249, 1999.
12. H. Mannila. Data mining: machine learning, statistics, and databases. In *Proceedings of the Eighth International Conference on Scientific and Statistical Database Management*, pages 1–8, Stockholm, June 18–20 1996.
13. H. Mannila. Local and Global Methods in Data Mining: Basic Techniques and Open Problems. In *Proceedings of ICALP 2002,* 29^{th} *International Colloquium on Automata, Languages, and Programming*, Malaga, Spain, July 2002. Springer.
14. H. Mannila. Theoretical Frameworks for Data Mining. *ACM SIGKDD Explorations*, 1(2):30–32, January 2000.
15. H. Mannila and H. Toivonen. Levelwise Search and Borders of Theories in Knowledge Discovery. *Data Mining and Knowledge Discovery*, 1(3):241–258, 1997.
16. T.M. Mitchell. *Version Spaces: An Approach to Concept Learning*. PhD thesis, Electrical Engineering Department, Stanford University, 1979.
17. T.M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.
18. L. Pitt and L. G. Valiant. Computational limitations on learning from examples. *Journal of the ACM*, 35(4):965–984, 1988.