Högdalenverket: Applying ILP in an Industrial Setting

Robert Engels *
Dept. of Computer and Systems Sciences
Stockholm University
Electrum 230, 164 40 Kista, Sweden

Abstract. When applying Inductive Logic Programming techniques in real-world settings, many problems will come up. The project was done in co-operation with Högdalenverket, a heat and power plant burning household refuse in the Stockholm area, Sweden. The application problems with collecting data and the application of ILP-techniques are discussed. Results of tests performed while using SPECTRE, an ILP-algorithm developed at Stockholm University, are reported. These results show that the addition of background knowledge and addition/retraction of parameters has positive effect on the performance of the ILP-techniques. After initial tests a knowledge acquisition stage was started. This resulted in knowledge about the domain that was used for evaluating the results of learning using SPECTRE. This prevented several mistakes by interpreting data and evaluating performance. This knowledge was partly used as background knowledge for the learning algorithm. Noise handling was applied and increased the efficiency of the SPECTRE-algorithm. The paper compares the performance of the SPECTRE-algorithm on theoretical databases with the performance on this practical domain.

Keywords : *Machine Learning, Inductive Logic Programming, Knowledge Acquisition*

1 Introduction

In the current paper we shortly[1] describe the problem of applying *Machine Learning* (ML) to a real-world domain. The domain is provided by Högdalenverket, a refuse burning plant. The aim of the project is to find control rules in order to lower the NO_x-emmision. *Inductive Logic Programming* (ILP) is used for learning in the domain. Few real-world applications using ILP are available for evaluation. The algorithm used for this application is called SPECTRE[2] and is developed at the Stockholm University [1]. Interviews at the plant and knowledge available from reports etc. were taken into consideration. The aim of the application of the ILP-techniques and the Knowledge Acquisition stage was to try to build and refine a model of the processes at the Högdalenverket power plant.

*The author's current address is: Institute AIFB, University of Karlsruhe, D-76128 Karlsruhe, Germany. E-mail: engels@aifb.uni-karlsruhe.de Tel: +49-721-608 4062 / Fax: +49-721-693 717

[1]Due to space constraints, a full version of this paper is published in the Proceedings of the Workshop on Intelligent Adaptive Systems, Melbourne Beach, Florida, 1995

[2]SPECialization by TRansformation and Elimination

2 Using ILP-techniques: the SPECTRE algorithm

The ILP-technique used was the SPECTRE algorithm [1]. This algorithm is based on the idea of finding an inductive hypothesis through the specialization of a logic program. The program has as input a logic program, a set of negative examples and a set of positive examples. Then the specialization problem (as commonly adopted in ILP) is defined as:

Given: a definite program P. Let E^+ be a set of positive examples and E^- be a set of negative examples such that $E^+ \cap E^- = \emptyset$ then *find:* a specialization of P (called P') such that $M_{P'} \subseteq M_P$ while $E^+ \subseteq M_{P'}$ and $E^- \cap M'_P = \emptyset$.[3]

The algorithm uses the measurement of impurity as defined in [12]. In SPECTRE an SLD-tree is pruned in order to search for a hypothesis. Boström and Idestam Almquist give a more extensive discussion on SPECTRE as well as the results of running SPECTRE on other domains [1]. In the current project these experiments will be extended by research involving the application of the SPECTRE algorithm on a real-world domain (See section 3).

3 The industrial setting: Högdalenverket

The domain which was evaluated consist of instances of the target clause T. Each clause has an initial 27 parameters. Each of these measurements involve the filtering and cleaning of emission products. The problem is provided by the Swedish Government; laws for the maximum-level of NO_x that emision gasses from the plant may contain are going to be restricted in the near future. Högdalenverket now couples the injection of NH_3 to one overall temperature measurement that is seen as representative for the temperature in the combustion chamber. That this is not an optimal solution is shown by the test-results which show the lowest accuracy when using this overall temperature measurement as the only parameter to learn from.

4 Symbolic Learning, Data transformation and Knowledge Acquisition

There are several reasons for applying ILP-techniques at the Högdalen domain. The Högdalen domain was selected because of the availability of relatively large amounts of data.

The Högdalen problem domain is also interesting because it deals with environmental protection and is therefore seen as a problem worth solving. This possibility to experiment with the application of new symbolic techniques provided a good opportunity to get more insight in the problems involved in applicating ILP in a real-world setting. At the same time there was a hope for some new insights in the processes involved in cleaning the emission-gasses. The specification of the problem that was defined at Högdalenverket was initiated by the fact that equipment at the plant was theoretically available for reducing of the NO_x-emission where in practice no control rules to do so were available. Due to the apparent possibility of improvement by simply finding control rules for the (already available) equipment and through the availability of already logged data it was seen as an attractive possibility to evaluate the SPECTRE algorithm in this domain.

[3] $M_{P'}$ denotes the least Hebrand model as described in [6]

The fact that the data were not represented purely symbolically, but numerical was problematic in the sense that few approaches are reported trying to integrate numerically and symbolical representation for use with ILP-techniques. Many practical domains however are represented numerically. Several authors recognize these and other kinds of (representational) problems (see for example [8], [2]).

The parameters in the logged data can be subject to noise. The parameters that cause the effects are not measured simultaneously. There always is a delay. The 5 minute averages are taken instead of the 10-second loggings while this approach filters out most of the delay effects. The data are corrected for incomplete and extreme data that are caused by measuring mistakes at the plant. The data are divided in subgroups according to standard statistical methods. These averages are then translated into semantically more meaningful groups.

Most ILP-systems can be used with different quantities and qualities of background knowledge. Adding background knowledge to the specialization problem has the advantage that the search process can be biased to raise the efficiency of the search process (see [11], [7], [5], [13], [3]). When databases grow larger this topic becomes of more interest. Background knowledge can be elicited performing Knowledge Acquisition and is normally represented in a symbolical form (This is one of the reasons for our approach to represent the example symbolical as well). The elicitation of Background Knowledge and its integration in ML projects are a topic that can only be mentioned here due to space constraints. As mentioned above adding background knowledge might increase efficiency and accuracy of the search algorithm by biased searching and avoiding (known) local maxima. Focussing is found a useful process in the stage of learning from databases ([7] for more on this topic). Interviews and reports resulted in a preliminary model of the processes involved and their connections. Partly this knowledge was selected to provide knowledge implemented in the form of intermediate predicates, were another part of this knowledge was used to focus the data. This kind of problem-solving is typical (in our view) for the application of new techniques in real-world applications.

5 Running SPECTRE on the Högdalen data

During a period of time tests with SPECTRE were run. The tests were performed in the same way as the evaluation of the SPECTRE-algorithm mentioned in [1].

The background knowledge that is collected is included in the testing. The tests evaluate the effects of the removing of literals and the addition of domain knowledge on accuracy. Redefinition of the general theory was done which included the addition of some intermediate predicates according to the gathered models.

It became clear that measuring only one overall temperature point (as done currently at the plant) is not enough to predict the amount of NH_3 that should be injected. Ignoring the HCl-measurements (after the gathering of domain knowledge it became clear that this literal could not be informative) provides us with a better performance for negative as well as positive performance. When adding the intermediate predicates (which provides extra grouping) the performance changes again. The performance is nearly as good as without adding intermediate predicates, but the learning produces less clauses (which can be seen as an improvement of performance, since few clauses means better coverage). This result makes it likely that adding more domain knowledge in this form could increase performance even more. Since we dealt with a real-world domain tests were run performing noise-handling by accepting clauses as they were

SPECTRE with:	Accuracy (%)	# clauses	Pos. acc. (%)	Neg. acc. (%)
All parameters	59.9	19.4	62.8	56.0
All parameters + 20 % noise	62.3	14.6	67.7	55.0
without HCl	61.1	19.5	63.1	58.5
without HCl + 20 % noise	62.0	14.9	65.1	58.4
Minimal set (temp only)	53.8	23.6	53.6	54.4
Minimal set (temp only) + 20 % noise	53.7	19.3	56.2	50.4

Table 1: Evaluating SPECTRE without noise handling and with a 20 % threshold (Datafile=0506/0507, No. of runs=100, NO_x-level=80 mg/nm^3, neg.ex. = 41 %)

covering more than 80 % of the positive examples and excluded more than 80 % of the negative examples. Some of the results of these tests are mentioned in table 1.

6 Conclusions

One of the reasons for applying SPECTRE in this particulair domain was that SPECTRE was previously tested and evaluated on several theoretical domains. A good reason for using inductive learning is that the output of the knowledge acquisition process will be in the form of models containing symbolic knowledge. The representation of learned knowledge will also be in a symbolic format. This means that the knowledge is easily interpreted by experts at the plant.

Although interviewing experts and processing reports does mean an extra time-consuming stage in a project, this can be useful and time-saving during the later stages of a project. Modelling background knowledge also results in a better understanding of the problem at hand. In our project the usage of background knowledge decreases the amount of clauses needed for the classication of the data-set. Only using one parameter as done now is certainly not enough for classification. The results of the second test series give an indication of which parameters should be taken into account.

There is a hope of finding results that might be transferred back to the plants technicians in the form of new, unknown concept-relation descriptions. This hope is based on reported successes of using inductive learning algorithms to find knowledge that is seen as new and interesting by experts and that triggered new knowledge (as reported in [4] and [9]).

Evaluating the project, it can be stated that using inductive learning algorithms on real-world domains without any knowledge acquisition process is not very helpful. As mentioned in [2] applying machine learning algorithms on "raw" data extracted from a database is of limited benefit. Our experience supports this point of view and we expect much benefit from the knowledge elicitation and representational redescription processes as described in [2]. They also stress the view on the development of systems as the one at hand as an iterative process.

The next stages of the project will include some further testing on newly provided data from Högdalenverket and testing the Högdalen domain with the MOBAL-system [10]. Representation of the learning problem as a problem of increasing/decreasing the ammount of ammonia injected at the several injection points is also planned. As mentioned before, it is also the intention to model and add more background knowledge in the course of the project. The combination of a knowledge acquisition stage and ILP has shown its value and will be needed in later stages of the project as well.

7 Acknowledgements

The work has been supported by the Swedish National Board for Technical and Industrial Development (NUTEK) as project "Machine Learning Techniques for Program Development" number 9303237/9405194, the Stockholm University and the University of Amsterdam. The help of Henke Boström with the application of his SPECTRE-algorithm and his comments on earlier drafts of this paper were invaluable.

References

[1] Boström, H. and Idestam-Almquist, P., "Specialization of Logic Programs by Pruning SLD-Trees", *Proceedings of the 4th International Workshop on Inductive Logic Programming*, volume 237 of *GMD-Studien, Gesellschaft für Mathematik und Datenverarbeitung MBH* (1994) 31–48

[2] Cupit, J. and Shadbolt, N., "Representational redescription within knowledge intensive data-mining." *Proceedings of the Third Japanese Knowledge Acquisition for Knowledge Based Systems Workshop*, Hatoyama, Japan, nov 7th – 9th, (1994)

[3] Feng, C., "Inducing Temporal Fault Diagnosis Rules from a Qualitative Model" In: Muggleton, F.H. *Inductive Logic Programming* (1992)

[4] King, R.D., Muggleton, S.H., Lewis, R.A. and Sternberg, M.J.E., Drug design by machine learning: The use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proceedings of the National Academy of Science USA*, volume 89, 11322–11326 (1992)

[5] Lindner, G., "Logikbasiertes Lernen in relationalen Datenbanken." (in German) *LS-8 Report 12*, University Dortmund (1994)

[6] Lloyd, J.W., "Foundations of Logic Programming." Springer-Verlag (1987)

[7] Matheus, C.J., Chan, P.K. and Piatetsky-Shapiro, G., "Systems for Knowledge Discovery in Databases." *IEEE Transactions on Knowledge and Data Engineering*, volume 5, number 6 (1993), 903–913

[8] Merckt, Th. van de, "Decision Trees in Numerical Attribute Spaces" *Proceedings of the 13th International Joint Conference on Artificial Intelligence, 28 august - 3 september*, volume 2, Chamberry; France (1993), 1016–1021

[9] Muggleton, S., King, R.D. and Sternberg, M.J.E., Protein Secondary Structure Prediction using logic-based machine learing. *Protein Engineering*, 5(4), 647-657 (1992)

[10] Morik, K., "Knowledgeable Learning using MOBAL- A Case Study on a Medical Domain." *Real-World Applications of Machine Learning; European Conference on Machine Learning*, Austria (1993)

[11] Muggleton, S. and Feng, C., "Efficient Induction of Logic Programs." *Proceedings of the First Conference on Algorithmic Learning Theory* Japan (1990)

[12] Quinlan, J., "Induction of Decision Trees." *Machine Learning*, volume 1 (1986) 81–106

[13] Sternberg, M.J.E., Lewis, R.A., King, R.D. and Muggleton, S., "Modelling the Structure and Function of Enzymes by Machine Learning" *Faraday Discuss.*, volume 93 (1992) 269–280

[14] Strömberg, A.M. and Karlsson, H.T., "Background Paper" presented to VärmeForsk Nordisc NO_x Seminar, Stockholm, march 17-19 (1993)