COIL 2000 Challenge Solution based on ILLM-SG Methodology

Tomislav Šmuc
Laboratory for Information Systems
Bijenička 54, 10000 Zagreb, Croatia
Phone: ++385-1-4561-085, Fax: ++385-1-4680-114
e-mail: smuc@rudjer.irb.hr
Rudjer Bošković Institute

ABSTRACT: This paper describes methodology used for solving tasks of COIL-2000 Challenge. Basic tools used in the detection task were ILLM algorithm (Inductive Learning by Logic Minimization) and stacked generalization. Induction of rules was performed in three steps. The first step involved optimization of parameters typical for ILLM induction process: noise level tolerance, maximum rule complexity. Optimized parameters were used in the final run and highest performance sub-rules were selected for stacked generalization. Description task involved explanation of sub-rules used in the detection task, and reconstructing 'customer models' from conjunction of attribute-value pairs.

KEYWORDS: COIL 2000 Challenge, inductive learning algorithm, stacked generalization.

INTRODUCTION

The problem of COIL 2000 Challenge represents well real-world problem. The dataset comprises actual data and tasks reproduce real requirements imposed on the data-miner: good prediction and explanation ability. Main difficulty of the dataset is its noisiness, which mostly requires special adjustments on the machine learning algorithms with respect to the objectives. This is the reason why in most of the communications between participants, other terms like lift-of or up-lift were used instead of accuracy. In the case of ILLM algorithm noisiness required a special, model optimization phase in order to obtain stable behaviour of induced rules. The form of induced rules by ILLM is conjunction of attribute-value pairs, and is rather suitable for the second task of the challenge.

TASKS

METHOD OF SOLUTION FOR THE FIRST TASK - DETECTION PROBLEM

COIL 2000 Challenge - detection problem was solved using ILLM (Inductive Learning by Logic Minimization) algorithm Gamberger (1995), and nearest neighbour stacked generalization algorithm (SG), which is still under development.
ILLM represents a propositional like machine learning algorithm. Its main features are: learning of CNF/DNF clauses, through minimization of number of attribute-value pairs (literals) necessary to satisfy learning samples. The objective function in the minimization problem is a complex function of noise level tolerance, complexity of the rule, and accuracy. In the problems like COIL–2000 detection problem one has to optimize parameters of the objective function in order to obtain more stable behaviour of induced rules (i.e. avoid overfitting).

The modeling process can be divided into three steps:

Model optimization

Optimization of the model includes determination of parameters typical for the ILLM algorithm such as: noise elimination level (very significant parameter in this problem), maximum number of sub-rules, maximum number of literals (attribute-value pairs) per sub-rule. These parameters were determined in a number of 5 fold tests, which involved different parameter combinations. This step showed that most stable rules (those having similar training and

test accuracy) have low complexity sub-rules and have to be induced with high noise level tolerance. 5 fold xv tests gave stable performance for several parameter combinations, and showed that in these cases training accuracy for the positive class over 20% of the dataset is 60-65%, and 50-55% for the test set. These combinations of model parameters were then used in rule induction over the whole training set, in the second step.

Determination and selection of best sub-rules

Determination of sub-rules for the detection was done using all training dataset samples and employing best (most stable) parameter combinations indicated in the first step. Those sub-rules having highest customer/no customer ratios were selected into a pool for final classification of the test dataset. It was enough to extract 11 sub-rules which were 'active' over slightly more than 20% of the training dataset, with average customer/no customer ratio of approximately 1/4.2.

When applied to the test dataset, sub-rules exhibited similar performance with respect to covering weight (% samples of the complete dataset) as on the training dataset, which was considered as a confirmation of stability of induced rules. This is indicated in Figure 1, which compares covering weight achieved on the training dataset against weight scored on the test dataset. Post-competition analysis shows that covering weight is just a weak confirmation of stability. This is illustrated in Figure 2, which shows up-lift of sub-rules on the training and test dataset. Figure 2 indicates overfitting, as all sub-rules have lower up-lift on the test dataset than on the training dataset.

Classification using sub-rules and stacked generalization algorithm

Stacked generalization (Wolpert, (1992)), is a layered learning technique for combining classifiers. The particular SG algorithm comprises a nearest neighbour algorithm at the second level, which in this case learns distinct sub-rule overlaps and associates class with it, based on some cost minimization rule. Its role is of greater importance in larger multiclass, cost sensitive problems. It was used here to combine individual sub-rules induced in the second step, and extract exactly 20% of 'most promising' samples from the test dataset as potential customers.

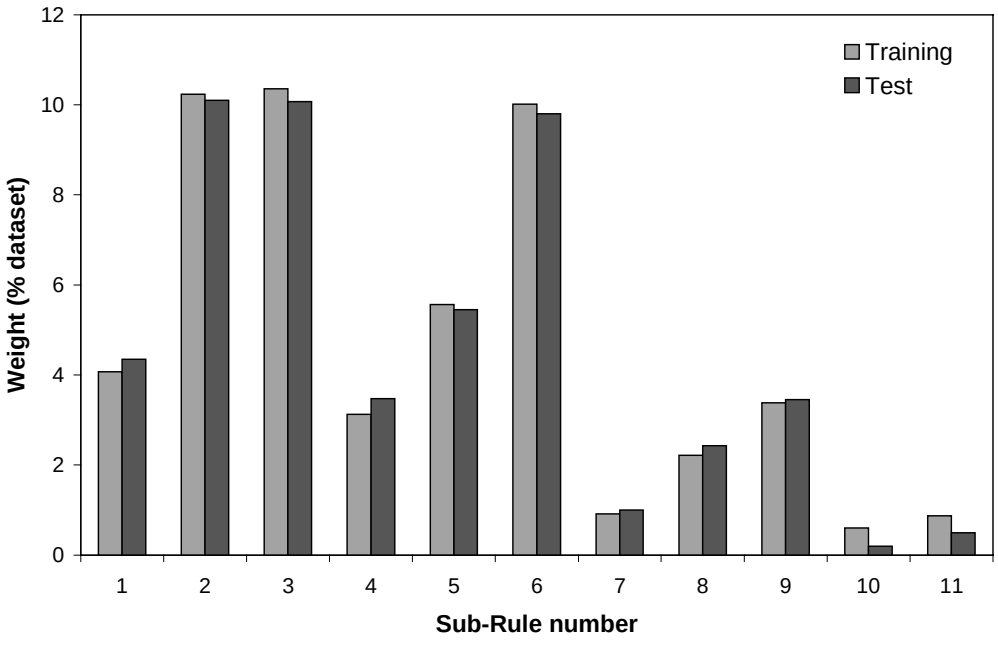

Figure 1: Covering weight of induced sub-rules.

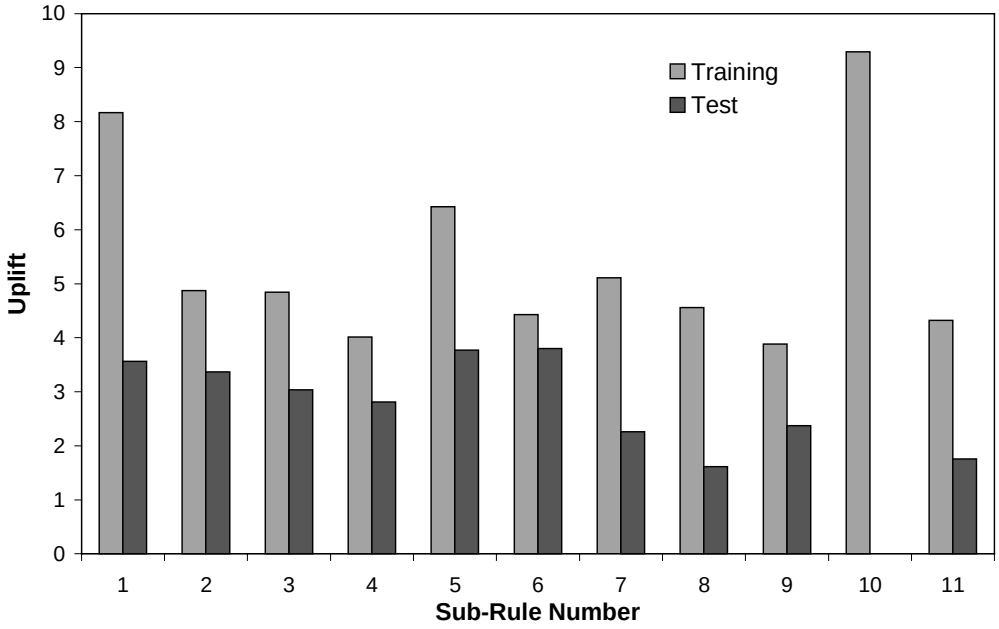

Figure 2: Up-lift of the induced sub-rules.

SECOND TASK - DESCRIPTION OF POTENTIAL CUSTOMERS

ILLM algorithm provides solutions in the form of conjunctions of attribute-value pairs. We call these conjunctions sub-rules, and they actually define subsets of samples (in this case potential customers). These sub-rules are informative and easily understandable even for non-professionals in machine learning. To have more generalization quality in the discussion, instead of using actual values, attribute-value pairs are described here using terms like *high*, *higher* or *lower than average* (average of the dataset).
In order to be even closer to non-specialists, I will give rough translation of the induced sub-rules used in the detection task, together with explanations (insight) of a layman in insurance policy business.

Majority customer group characteristics

Potential customers are described by sub-rules defined through conjunctions of attribute-value pairs. The highest hit-rate have the sub-rules containing following two attribute-value pairs:

High contribution to fire policies: (PBRAND = 4 or 3)
High contribution to car policies: (PPERSAUT = 6)

These two (in a number of sub-rules also together), indicate most probable owners of caravans. The high contribution to car policies is probably dictated by the insurance companies' rules (additional risk for the car), and probably by the fact that people owning caravans have larger cars. High contribution to fire policies reflects greater risk associated with fires in caravans. Additional attribute-value pairs which define sub-rules in conjunction with one or both of these two most important attribute-value pairs, indicate more specific characteristics of most probable customers:

Customer subtypes are generally younger, or stable, middle class families with children (customer subtype 8-13), living in the areas with:

a) Higher than average Average Income
b) Average and higher than average number of people with Income 45-75.000;
c) No farmers
d) Average and lower than average Low level educated people
e) Average and lower than average Number of singles
f) Average and higher than average Purchasing power class
g) Average and higher than average Married people
h) Average and lower than average No religion people
i) Areas with average and higher than average Number of Protestants
j) Higher than average Middle management people
k) Lower than average Social Class B2 people

Altogether, there were 6 distinct sub-rules defined by one or both of the most important attribute-value pairs (PBRAND and PPERSAUT) in combination with one or more from the a)-k) list. These defined the majority of most probable customers in extracted model. The subsets of potential customers defined by these sub-rules are characterized by rather high degree of overlapping, which is why the list of all used attribute-value pairs is given. From this list insurance professional, knowing well other population characteristics, could distinguish even more logical subsets than those defined by the dataset and indicated by the machine learning algorithm. For a layman in the insurance policy business, this overlapping only indicates rather broad characteristics of most probable customer.

Distinct 'minority' customer groups' characteristics

The modeling also indicated sub-rules involving other attribute-value pairs which define distinct, "less numerous" potential customer groups, but with rather high customer/no-customer ratios:

- People having boat policies, living in the areas with more than average number of rented houses. (Explanation: This might indicate holiday-fanatics, living most of the free time in tourist-seaside areas, possibly already owners of caravans, or at least future potential customers);

- Customer subtypes in the class of elderly (50+) people, or larger families, from small towns, living in the areas with more than average Social Class D people, and no farmers. (Explanation: Social Class D may be also indicates elderly (probably knowledge of this attribute would change reasoning that follows?). These are may be 'garden owners', from urban and sub-urban areas, or even poorer people living in caravans);

- People living in areas with rather high level of Protestants, with certain number of people in high income range (75-122.000), and lower than average number of people with lowest income range, with higher than average NHS. (Explanation: Not a very well defined customer group. Some of the experiments indicated that this might be town dwellers, of middle management, medium education type, of a rather high purchasing power class, motivated for caravans probably by children or globe-trotting inclinations);

- People not having fire policies in the areas with rather low NHS, low number of low educated people, average and higher than average number of skilled labourers, and no social Class D people. (Explanation: Again not a group with sharp contours. Probably indicates people from younger, working and small entrepreneur class areas (NHS?), with moderate and higher income. It could be that they are very similar to the group above.)

These subsets, in contrast to first, majority group, may not have high contributions for car and fire policies, but do have the mobile home policy. Probably due to one of the following reasons:

- their mobile homes are not practically mobile, i.e. they don't need car policies;
- their fire policies are included in mobile home policy;
- they don't have money for all these policies, so they cover only the most important ones, or their contributions are "more reasonable";
- they are fresh mobile home owners; i.e. they didn't have time to get all the policies they want.

I hope that this discussion gives some clues to insurance experts, and provides some arguments for the existence of our (AI, ML, EC) community.

ACKNOWLEDGMENTS

I would like to express gratitude to my colleague Dr. Dragan Gamberger who is the author of ILLM code and who has helped me with advises. I also want to express my gratitude to the organizers for giving us the opportunity to work and learn (in good company and atmosphere!) on a well-defined real world-problem.

REFERENCES

Gamberger, Dragan, 1995, “A Minimization Approach to Propositional Inductive Learning”, . In Proc. 8th European Conference on Machine Learning (ECML-95), Berlin, pp. 151-160.

Wolpert, D., 1992, “Stacked Generalization”, Neural Networks 5, pp. 241-259.