Solution based on ILLM confirmation rule

Dragan Gamberger
Rudjer Bošković Institute
Bijenička 54, Zagreb, Croatia
e-mail dragan.gamberger@irb.hr

METHODOLOGY

The submitted solution was induced by ILLM (Inductive Learning by Logic Minimization) system. It is an Occam's razor based inductive system which presents knowledge in form of rules. The main building blocks for the rules are literals representing simple logical conditions of domain attributes. This knowledge representation form enables easy and comprehensible interpretation of induced results what was very important for the Coil challenge. The ILLM system can, as one of its options, generate rules in the so called confirmation rule set form. This form is typically used for applications requiring reliable target class prediction. The form consists of a set of confirmation rules so that every confirmation rule must be a simple conjunction of literals, every rule must be true for examples of the target class only, and every confirmation rule must be true for at least predefined number of target class examples defined by the selectable support level. Number of induced rules in the set is determined by the support level so that all confirmation rules satisfying this level are included into the set. Typically the user determines to include only few confirmation rules into the final set. In this case selection among all acceptable confirmation rules is based on their covering properties for the target class examples. In case of Coil challenge, confirmation rule set form is selected because such rule form is especially easy for human interpretation. This requirement restricted also the total number of induced rules in the set. The final solution consisted of only one confirmation rule. The main problem of confirmation rule induction in this domain was very high level of noise among training data. Typically the problem of noise in ILLM is solved in preprocessing by an explicit noise detection algorithm. In this approach, detected noise is eliminated from the training set before the rule induction process. In the Coil domain, additionally, it was necessary to allow that some non-target examples are covered by the induced rules as well. By selecting the optimal ration of the number of covered target class examples and the number of covered non-target class examples it was possible to induce rules covering about 20% of all examples, what was the condition in this competition.

RESULTS

The induced confirmation rule has the following form:

$$(a05 \neq 10)\ (a11 \leq 2)\ (a12 \leq 4)\ (a18 \leq 5)\ (a37 \leq 4)\ (a43 \geq 2)\ (a47 \geq 6)$$

Based on the rule as a good prediction model for caravan policies customers we suggest a simple model consisting of only 7 conditions. These seven conditions must ALL be satisfied in order that the model describes a potential policy customer. The conditions are:

Customer main type is	not farmers	5 MOSHOOFD	is not 10 (farmers)
Living together number is	low	11 MRELSA	is 2 or less
Other relation number is	low	12 MRELOV	is 4 or less
Low level education number is	low	18 MOPLLAAG	is 5 or less
Income < 30.000 is	low	37 MINKM30	is 4 or less
Purchasing power class is	high	43 MKOOPKLA	is 2 or more
Contribution car policies	high	47 PPERSAUT	is 6 or more

The conditions demonstrate importance of selected attributes and define limits for their acceptable values. For example, both 'living together' and 'other relation' descriptor values must be in this model low, but for 'living together' acceptable values are only 0, 1, and 2, while for 'other relation' descriptor acceptable are all values between 0 and 4. It this sense, it is both important and informative which descriptors are used in the model and which values are selected as acceptable.

Short description of the caravan policies customer based on this model is:
'customer main type' is not farmers AND in his living area there are very low number of 'living together' people AND low number of 'other relation' people AND low number 'low level education' people AND low number of 'income < 30.000' people. Also in the area there must be many 'purchasing power class' people AND the customer's contribution of car policies' must very high. The model suggests that potential custemers must be searched in non-rural communities with low number of low income people, low number of people living in strange family relations, and low number of low education people. In such communities potential customers must be searched among purchasing power class people with high number of car policies.

This model is restricted in the sense that it describes about only 50% of potential customers. Simplicity, comprehensibility, and usefulness are its advantages.

APOSTERIOR RESULT ANALYSIS

The induced model correctly predicted 105 (44%) customers in the test set. The result is below the estimated value of 50% but the relative small difference shows that the induced rule is not overfitted and that the noise handling approach was appropriate. The main value of the prediction result is in the fact that it is obtained by only one, relative simple rule. The description quality of the induced knowledge is that potential customers are represented by a conjunction of 7 simple conditions and that every of these conditions is completely justified and comprehensible. In this model nor subconcepts could be detected but it represents a high quality general description of potential customers. The 6 out of 7 conditions in this model are demographic data and information contained in them can be directly used to locate communities with unproportionally high number of potential customers.

REFERENCES

N. Lavrač, D. Gamberger, P. Turney, 1998, " A relevancy filter for constructive induction". *IEEE Intelligent Systems & Their Applications*, **13**:50-56.

D. Gamberger, N. Lavrač, S. Džeroski, 2000, " Noise detection and elimination in data preprocessing: experiments in medical domains". *Applied Artificial Intelligence*, **14**:205-223.

.