COIL Challenge 2000

Petr Mikšovský, Jiří Klema

Czech Technical University,
Prague, Czech Republic
{miksovsp, klema}@labe.felk.cvut.cz

Dataset

The training dataset is unbalanced. There is only about 6% of positive examples. Moreover, some of them (about 50) are contradictory, i.e. exactly the same combinations of attribute values appear in negative examples as well.

The original dataset was cleaned in the way that all contradictory records were re-classified as positive. In the next step there were derived additional attributes suggested during the Sol-Eu-Net workshop in April 2000. The first part of the new attributes corresponds to transition from a many-valued attribute to a binary one. This was applied to attributes 44 to 85 in the following way: if a value of the considered original attribute is greater then 0 than assign 1, otherwise 0. In the second part there are introduced two new attributes *Sum_of_contributions* and *Sum_of_policies* defined as a sum of values corresponding to *contribution* (attributes no. 44-64), resp. *policies* (attributes no. 65-85) attributes. The final part contains derived attributes aggregating all policies of the considered customer by the purpose (i.e. vehicles, third party, etc.).

Processing and description of interesting customers

The extended dataset was processed by number of tools, e.g. KEPLER, C5.0, 4FT-MINER, and iBARET. For the description of the last two ones see the next paragraph. We have succeeded to find some rules, which seem to be very reliable in identifying negative examples (i.e. those customers who do not buy a caravan policy), e.g.:

1. If *number of moped policies* is greater than 0 and *number of car policies* is 0 then the class is NO (100% accuracy).
2. If *number of houses* is greater than 2 then the class is NO (100% accuracy).
3. If *income > 123.000* is greater than 3 then the class is NO (100% accuracy).

The same extended dataset was used by iBARET to learn the concept in the form of a model described by a vector of weight values. The highest weights were obtained for the following attributes (ordered by importance): 47 (*Contribution to car policies*), 59 (*Contribution to fire policies*), 80 (*Number of fire policies*), *Sum_of_contributions*, 40 (*Income 75-122.000*), *Sum_of_policies*.

The classification of the challenge dataset proceeded as follows: the sure negative examples were eliminated using the rules mentioned above. This led to a simplified dataset consisting of 3800 unknown records. For the rest, the iBARET model was applied and suggested for each record the "estimate of its probability of being positive". The examples with the highest estimate were considered as the most promising ones. The first 800 of those examples were submited.

Methods used

iBARET (Instance-Based REasoning Tool) is an implementation of the principal characteristics of IBR theory. It represents a universal tool for modeling and predicting in domains described by a number of numeric or symbolic values with restricted or no background knowledge. It consists of two distinct parts:

- Database complemented by a search engine, which searches a case memory for the nearest neighbors,
- IBR Modeling Interface, which generates queries to the database, evaluates its responses and adjusts feature weights.

iBARET represents a batch feature weighting method with performance bias, i.e. it uses feedback from the performance function during training. This function is calculated for each setting of feature weights and a genetic algorithm is used to suggest new sets of feature weights. Receiver operating curve methodology is used to calculate this fitness function, the methodology is able to optimize classifier performance in domains with non-uniform class distribution.

4FT-Miner (GUHA procedure) deals with generalized association rules of the form Ant ~ Suc/Cond, where Ant, Suc and Cond are complex derived Boolean attributes (e.g. conjunctions of disjunctions). Symbol ~ is a 4FT quantifier and expresses a dependency between Ant and Suc. Various classes of dependencies can be used, e.g. implications (this corresponds e.g. to the "standard" association rules evaluated using support and confidence), double implications or symmetrical (e.g. chi^2, Fisher or equivalence).

Acknowledgement

This work was supported by the Sol-Eu-Net project IST-1999-11.495.

Reference

[1] Klema, J. - Lhotska, L. - Palous, J. - Stepankova, O.: Instance-Based Modeling in Medical Systems. Cybernetics and Systems 2000, Proceedings of 15th European Meeting on Cybernetics and Systems Research (EMCSR), (R. Trappl, ed.), Vienna, 2000, Austrian Society for Cybernetic Studies, ISBN 3 85206 151 2, pp. 365-370.

[2] Berka,P. - Rauch,J.: Data Mining using GUHA and KEX. In: (Callaos, Yang, Aguilar eds.) 4th. Int. Conf. on Information Systems, Analysis and Synthesis ISAS'98, 1998, Vol 2, 238- 244.