Chapter 3

Extended Abstracts

Automated Knowledge Acquisition for PROSPECTOR-like Expert Systems

Petr Berka, Jiří Ivánek

Dept. of Information and Knowledge Engineering, Prague University of Economics,
W. Churchill Sq. 4, 130 67 Prague, CR

Abstract. The method for automatic knowledge acquisition from categorical data is explained. Empirical implications are generated from data according to their frequencies. Only those of them are inserted to created knowledge base whose validity in data statistically significantly differs from the weight composed by the PROSPECTOR like inference mechanism from the weights of the implications already present in the base. A comparison with classical machine learning algorithms is discussed. The method is implemented as a part of the Knowledge EXplorer system.

1 Knowledge Acquisition task

The aim of an particular application of a diagnostic expert system is to weight each diagnosis (goal of the consultation) using values of input attributes. A knowledge base of such a system contains for each goal a set of weighted rules leading from combinations of values of input attributes to this goal.

In our sence, PROSPECTOR-like expert systems are based on rules in the form

$$Ant \Longrightarrow Suc \ (weight)$$

where

$Ant = j_1 c_1 ... j_k c_k$ is a combination (conjunction of attribute-value pairs, also called *categories*) of length k,

Suc is a single category (goal),

weight from the interval $< 0, 1 >$ expresses the uncertainty of the rule.

During a consultation, all the rules which match the values of input attributes of a particular case are found and their weights are combined (composed) using following pseudobayesian combining function $\oplus$: $x \oplus y = (x * y)/(x * y + (1-x) * (1-y))$. Since the knowledge base can contain both a rule $Ant \Longrightarrow Suc \ (weight)$ and its subrule $Ant' \Longrightarrow Suc \ (weight')$, where $Ant' \subset Ant$, this operation is used with respect to the correction principle suggested by Hájek [6].

The result of a consultation is a list of goals (diagnoses, recommendations, etc.) ordered according to their composed weights.

Our idea of *knowledge acquisition* is to construct the knowledge base as a minimal set of rules, which describes given data and can directly be used for consultations. The essential question is which of empirical implications $Ant \Longrightarrow Suc$ (where Suc is the goal diagnosis) hidden in data are to be inserted into the resulting knowledge base and with which weight. The answer depends not only

on the data (i.e. on the validity of each implication computed as conditional probability $P(Suc/Ant)$) but also on the inference mechanism used (i.e. in our case that of PROSPECTOR), and on the requirement of accuracy with which the resulting knowledge base corresponds to the data. The predictive power of the knowledge base is controlled by the employed statistical test of the hypotheses that results obtained from knowledge base during consultations (expressed as composed weights) correspond to empirical implications in data (expressed as validities).

2 Algorithm

Input: Data, goal Suc
Output: Knowledge base KB

Initialisation:

Set KB to be a list containing the empty implication $\emptyset \Longrightarrow Suc$ with the weight computed from the relative frequency of Suc in data;
Set CAT to be a list of categories jc sorted in descending order of their frequencies in data;
Set OPEN to be a list of implications $jc \Longrightarrow Suc$ sorted in descending order according to the frequencies of jc in data ;

Computation:

while OPEN is not empty do

1 select the first implication $Ant \Longrightarrow Suc$ from OPEN;
2 test of the implication $Ant \Longrightarrow Suc$:
  2.1 compute the *validity* of the implication;
  2.2 compute the *composed_weight* from the weights of all the subrules of $Ant \Longrightarrow Suc$ which are already in KB, using composition function $\oplus$;
  2.3 if the validity significantly differs from the composed weight (we use the χ^2 goodness-of-fit test) then add $Ant \Longrightarrow Suc$ to KB with the weight w such that $w \oplus$ *composed_weight* = *validity*;
3 expansion of the implication $Ant \Longrightarrow Suc$:
  3.1 for each jc from CAT such that jc precedes in CAT all the categories from Ant do
    3.1.1 generate a new combination $Ant \,\&\, jc$;
    3.1.2 insert the implication $Ant \,\&\, jc \Longrightarrow Suc$ into OPEN just after the last implication $C \Longrightarrow Suc$ such that the frequency of C is greater or equal than the frequency of $Ant \,\&\, jc$;
  enddo;
4 delete the implication $Ant \Longrightarrow Suc$ from OPEN;

enddo;

If required, the expansion of implications (Step 3) can be controlled by a required maximal length of Ant, a minimal required frequency of Ant and a minimal required validity of $Ant \Longrightarrow Suc$.

3 Experiments

The presented algorithm of knowledge base acquisition is implemented as a part of the Knowledge EXplorer system. [1, 7]. The knowledge acquisition component has been tested in the framework of the ALEX system [10] and for evaluation of virological hepatitis tests [2].

We also compared our approach with CN2 and KnowlewdgeSeeker (KS)[1] on data taken from Machine Learning Repository [8]. "Japanese Credit data" consists of 125 objects and 11 attributes; the task is to learn knowledge when to grant a credit. The whole set was used for training. "Monk's data" are three artificial problems used for testing different ML algorithms [9]. The results are summarized below. The table gives number of the obtained rules and the accuracy of classification (on training set for data CREDIT, on testing sets for data MONK):

	data CREDIT		data MONK1		data MONK2		data MONK3	
system	rules	accuracy	rules	accuracy	rules	accuracy	rules	accuracy
CN2	35	100%	8	100%	51	66%	19	91%
KS	15	80%	3	75%	1	67%	12	94%
KEX	86	97%	4	50%	59	66%	10	99%

The accuracy 50% for KEX on MONK1 data is caused by missing rules for negative examples; so these examples remained unclassified. On the other hand, all positive examples were classified correctly. In this case KEX learned exactly the hidden concept.

4 Concluding remarks

Knowledge EXplorer performs symbolic empirical learning from examples (cases), where the induced concept description is in the form of weighted decision rules. Our algorithm can deal with noisy data, unknown values, redundancy and contradictions.

The generalisation (done by selecting implications using the χ^2 test and removing of redundant rules) is usually very high. Typically, the resulting knowledge base consists only of a small fraction (several percents) of all implications that fulfil the parameters.

When visually interpreting the knowledge base, sometimes some "obvious" piece of knowledge cannot be found. This is because the effect of the corresponding "missing" rule can be composed from its (more general) subrules, which are already in the knowledge base. So this rule is redundant and thus not inserted. Therefore, the knowledge base has to be taken into account as a whole and only within an expert system.

[1] Commercial TDIDT system developed at FirstMark Technologies, Canada [3].

When comparing Knowledge EXplorer to well known learning algorithms (TDIDT-like and AQ-like) we can find the following differences:

1. KEX can create more than one rule covering a specific example,
2. the knowledge base of KEX can contain both a rule and its subrule,
3. during consultation, the system can recommend (infer with positive weight) more than one concept.
4. because of used statistical test, KEX requires reasonable amount of input data,

The resulting set of rules obtained by KEX is usually larger then that obtained by ID3-like or AQ-like algorithms. This fact gives more possibilities for explanation and allows us to handle incompletely described cases since more then one rule is applicable during consultation. But the obtained knowledge base is closely related to PROSPECTOR-like inference mechanism.

References

1. Berka, P.: Knowledge EXplorer. A Tool for Automated Knowledge Acquisition from Data. TR-93-03, OeFAI Technical Report, Vienna, 1993.
2. Berka, P.: A Comparison of Three Different Methods for Acquiring Knowledge about Virological Hepatitis Tests. TR-93-10, OeFAI Technical Report, Vienna, 1993.
3. Biggs,D. - de Ville,B - Suen,E.: A method of choosing multiway partitions for classification and decision trees. Journal of Applied Statistics, Vol. 18, No. 1, 1991, 49-62.
4. Clark,P.: Functional Specification of CN and AQ. TI/P2154/PC/4/1.2, Turing Institute, 1989.
5. Duda,R.O. - Gasching,J.E.: Model Design in the Prospector Consultant System for Mineral Exploration. in: Michie,D. (ed.), Expert Systems in the Micro Electronic Age, Edinburgh University Press, UK, 1979.
6. Hájek,P.: Combining Functions for Certainty Factors in Consulting Systems. Int.J. Man-Machine Studies 22,1985, 59-76.
7. Ivánek,J. - Stejskal,B.: Automatic Acquisition of Knowledge Base from Data without Expert: ESOD (Expert System from Observational Data), in: Proc. COMPSTAT'88 Copenhagen (Physica-Verlag Heildelberg 1988), 175-180.
8. Murphy,P.M. - Aha,D.W.: UCI Repository of Machine Learning Databases. Irvine, University of California, Dept. of Information and Computer Science.
9. Thrun S.B. et al: The Monk's problems. A Performance Comparision of Different Learning Algorithms, Carnegie Mellon University 1991, 154p.
10. Winkelbauer,L. - Berka,P.: New Algorithms for ALEX: Expanding An Integrated Learning Environment. in: Proc. ECML93 workshop on integrated learning architecture, 1993.