Applying Decision Trees in Classification Tasks

Galant Violetta
University of Economics
Wroclaw, Poland
galant@ksk-2.iie.ae.wroc.pl

Owoc Mieczyslaw L.
University of Economics
Wroclaw, Poland
owoc@ksk-2.iie.ae.wroc.pl

Gladysz Tomasz
University of Economics
Wroclaw, Poland
gladysz@manager.ae.wroc.pl

Abstract

From its nature, decision-making processes and classification tasks are domains, where decision trees (DTs) are widely applied. Power of DTs to represent some knowledge structures seems to be obvious. In spite of, decision trees have been introduced as an universal tool - their properties and range of applications look to be limited to some preferred tasks. The paper demonstrates the utility of the mentioned technique in modelling of chosen knowledgebases, especially in certain classification tasks. An overview of classification algorithms in the context of generating decision tree precedes the more practical considerations. The problem of evaluation of bank customer's creditability in a face of crediting is put as an example. The real testing databases are taken into account as the learning files. Some suggestions on using DTs in similar tasks are presented in conclusion.

1. Introduction

Basically, expert systems are created for supporting different activities of decision-making processes. Therefore these system can be used for classification tasks, diagnosing, monitoring, configuring, planning and the like purposes (see for example: [5]). Classification tasks seem to be rather simple, that why – a quite significant number of applications represent this goal [2]. On the other hand, these tasks are employed as supporting procedure for such systems like diagnosing, prediction or interpretation.

The technique can be regarded in the paper is a decision tree. A decision tree has been introduced as an universal tool supporting mostly a problem's search space.

The classical tree consists of two main parts: nodes and arcs, which link related nodes. Nodes represent chosen decision issues. The arcs are used for expressing possible values for each issue. In such light, it is very easy to create a tree, which can be used for classification purposes.

The content of the paper is as follows. The next section presents the issues of classification tasks. Therefore, examples of this sort of tasks are pointed out. The topic of the next part is to characterise classification algorithms, which can be used for generation of decision trees. Short notes referring to the itemised procedures are included. The next section describes applying of decision trees during classification of customers, which put credit applications in a bank. The last part comprises of conclusions, derived from the research.

2. Classification Tasks

The classification tasks are regarded very important, especially in the context of expert systems. The classification task is everywhere, where we are choosing one decision from many possible. The exemplars of classification tasks are:

- credit decision making based on client's financial situation,
- specific management decision making from economic ratios,
- estimation of appropriate post and salary for applying candidate,
- definition of the best client for given group of goods in marketing research.

They represent different sorts of possible tasks, however this is an open list.

A formal definition of the classification task is:

Objects used for classification knowledge generation are called the examples and are given in a training set C. The examples in training set describe m attributes X and one classification attribute Y. Each example in the C set describes an entity as follows [8]:

$$C_i = (x_1, x_2, \ldots, x_m, y)$$

$$where: x_l \in dom(X_l), y \in dom(Y), l = 1, \ldots, m$$

On the basis of a training set C, the rule of a classification set φ is generated such as:

Proceedings of the Workshop on Computer Science and Information Technologies CSIT'99
Moscow, Russia, 1999

$$\phi(x_1,\ldots,x_m) = y$$

$$where\colon x_l \in dom(X_l), y \in dom(Y), l = 1,\ldots,m$$

The result of the system solving classification task is knowledge allowing to prescribe new examples (not belonging to a training set C) to one of determined classes. This effect can be presented by means of many knowledge representation formalisms, like: production rules, semantic networks or frames.

3. Overview Classification Algorithms for Decision Trees Generation

The form of such tree is constructed from the root to the leaves. Hunt [6] introduced the result of the classification problem in form of decision tree. The Hunt's algorithm for the tree creating was called CLS (Concept Learning System).

Many people on the basis of this algorithm have proposed their own ways of developing decision trees. Selection of the most convenient measures influences the size, comprehensibility of decision tree and accuracy of classification. The decision tree should be formed in such a way that the tree classification was done very rapidly (that means the smallest possible number of attributes should be used) on condition that the classification accuracy is preserved.

Algorithms CART and C5.0 belong to the most popular decision tree learning system.

- **CART**

It was described by Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. [1]. In this algorithms was applied the function, which measures the impurity of nods in decision tree. The impurity of this node is maximal, if the classification attribute is stochastically independent of described attribute. But the impurity would be completely removed when the attribute is found in the node and is functional dependent of classification attribute. In CART this intuitive idea of impurity is formalised in the GINI index for the current node *c* [7]:

$$\boldsymbol{GINI}(c) = 1 - \sum_{j=1}^{k} p_j^2,$$

where p_j is the probability of class j in node *c*.

For each possible split the impurity of the subgroups is summed up and the split with the maximum reduction in impurity chosen.

- **C5.0**

J. R. Quinlan developed CLS algorithm proposed an evaluation function based on a classic formula from information theory that measures the theoretical information content of a code [9, 10]:

$$E = -\sum_{i=1}^{n} p_i \log(p_i),$$

where pj is the probability of i-th message

On the base of entropy J. R. Quinlan defined the gain criterion. It measures the increment of information for choosing given attribute in the node. In the last version of Quinlan's algorithm C5.0 was applied adaptive boosting, based on the work of R. Shapire and Y. Freund. The idea is to generate several classifiers. One of the most important features, which C5.0 incorporates, is variable misclassification cost.

The authors of this paper created own decision tree learning system - GIMS (Generalisation by Inductive Symbolic Method) – [3, 4], which was applied to generate decision trees from a banking domain. In the GIMS system, in order to generate a decision tree, Czerwinski coefficient of attributes association was applied. This coefficient measures degree of dependence, or independence, which exists between two variables (Galant, 1996). The coefficient is computed for all described attributes. The maximum value of coefficient decides on the choice of the attributes to the following nodes of the decision tree.

4. Applying Decision Trees Algorithm

Credit decision making, based on client's financial condition, is an example of the classification task. The tests were based on two databases. The first one - CREDIT_P concerns consumption credits, but the second one – CREDIT_E includes dates for economic activity. The both databases contain the real data files from Polish banks. Table 1 describes briefly main features of the training sets.

Table 1. Test Databases characteristics

	No. of	No. of	Attributes	
Database	cases	classes	continuous	discrete
CREDIT_E	125	2	5	5
CREDIT_P	146	2	1	5

CREDIT_P

Three classes of people which took consumption credit was defined:

- class 1 – **yes** – if credit was gave back in a time,
- class 2 – **time** – if the credit was gave back not in a time,
- class 3 – **no** – if the credit was not gave back.

Customers were described by means of six numeric attributes and six symbolic attributes. In the research it was taken into account following numeric attributes:

- **value** - credit total,
- **part** – part payment,
- **age** – age of people,

- **income** – month incomes,
- **expense** – month expenses,
- **person** – number of person in family.

Also it was taken into account following symbolic attributes (with set of values):

- **credit (car, something, cash) –** kind of credit,
- **sex (f, m)** – sex of people,
- **place (city, town, village) -** place, where live people,
- **source** (worker, retired, other) – source of income,
- **guarantee** (pawn, surety, other, two) – sort of guarantee,
- **status (couple, single) –** civil status.

CREDIT_E

In this learning set it was analysed one numeric attribute (**Financial Ratio**) and five symbolic attributes:

- **sale** – possibility of sale,
- **forecast** - sales forecast,
- **management** – estimate of management,
- **guarantee** – level of guarantee,
- **demand** – market require.

It was generated decision trees for each databases.

```
DECISION  TREE  CREDIT_P

----------------------------------------------------------------------

VALUE - VALUE<550.00  --> yes
VALUE - VALUE>550.00
        VALUE - VALUE<25500.00
        |    SOURCE - WORKER
        |    |   INCOME - INCOME<540.00  --> time
        |    |   INCOME - INCOME>540.00
        |    |       INCOME - INCOME<913.50  --> no
        |    |       INCOME - INCOME>913.50  --> yes
        |    SOURCE - RETIRED
        |    |    INCOME - INCOME<1118.00  --> no
        |    |    INCOME - INCOME>1118.00  --> yes
        |    SOURCE - OTHER
                |            |            KREDYT - SOTHING  --> yes
        |    |  KREDYT – CAR      --> yes
        |    |  KREDYT – CASH     --> no
        VALUE - VALUE>25500.00  --> yes
```

Figure 1. Decision Tree for CREDIT_P

The analysis of this decision tree from Figure 1 gives us the same direction. Mainly very small (< 550) and very big (>25500) credits was repaid. If credit customer is a worker and his month income is not high than 540 it would be a need to remember him about the required istalment. If his month incomes were between 540 and 913,50 the credit would not pay off. Beyond this partition probable the credit would pay off. Similarly, it is if retired took the credit. In this situation the risk of credit depends of an income level. In the case other source of income credit risk depends of kind of credit. The cash credit has the highest risk.

This decision tree from Figure 2 shows us that the most important attribute was "demand". If credit customer produced goods with high market require, they usually repaid the credit. On the contrary the low level of "demand" marked that credit often was not pay off. In the case, when "demand" were good and average attributes: "Financial ratio" and "Sales Forecasts" had the decisive role.

```
DECISION    TREE  CREDIT_E
-------------------------------------------------
    Demand - high  --> yes
    Demand - good
    |        Financial ratio<=21,30 --> no
    |        Financial ratio>21,30 --> yes
   Demand = average
    |        Sales Forecasts = v.good → yes
    |        Sales Forecasts = good
    |         |            Financial ratio <=17.50 → no
    |            |         Financial ratio >17.50 → yes
    |            Sales Forecasts = sufficient → no
    |            Sales Forecasts = insufficient → no
   Demand = low → no
```

Figure 2. Decision Tree for CREDIT_E

5. Conclusion

The analysis of the decision trees, which are generated on the base of credit history, give us same directions about the definite group of credit customers. It can be useful by preparing credit instructions. DTs can be applied in similar tasks, where the need of classification appears and the list of classification attributes is limited to low numbers..

Generally speaking, decision trees allow for classifying the new bank customers. The decision tree system prompt only but the human expert makes a last decision. But, what is important, the proposal of final decision is generated by the system.

This way, usability of DTs in classification task was proved. In practice, when the learning set is not properly prepared the final decision tree would be wrong. The same inadequate effects can be achieved, when the attribute list contains too many positions or the value list is composed of too heterogenous items.

References

1. Breiman l., Friedman J.H., Olshen R.A., Stone C.J.: Classification and Regression Trees. Wadsworth and Brooks 1984.
2. Durkin J. (1993): Expert Systems: Catalogof Applications, Intelligent Computer Systems
3. Galant V. (1996), GIMS - Decision Tree Learning System, in: Proceed. of the 1st Polish Conference on Theory and Applications of Artificial Intelligence, Lodz.

4. Galant V. (1997), Zastosowanie indukcyjnych metod symbolicznych do odkrywania wiedzy w SIZ. [Application of Inductive Symbolic Methods for Knowledge Discovery in MIS]. Doctoral Dissertation, Wroc³aw (in Polish)

5. Hayes-Roth F., Waterman D.A., Lenat D.B.: Building Expert System, Addison-Wesley, Reading, Mass.

6. Hunt E.B., Marin J., Stone J.P.: Experiments in Induction. Academic Press 1966.

7. Michie D., Spiegelhalter D. J., Taylor C.C.: Machine Learning, Neural and Statistical Classification. Ellis Horwood Limited 1994,

8. Owoc M.L., Galant V.: Validation of Rule-Based Systems Generated by Classification Algorithms, in: Proceed. of the Information Systems Development Conference, Bled'98 – Slovenia (to appear)

9. Quinlan J.R.: Induction of Decision Trees. Machine Learning 1986/1

10. Quinlan J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers, 1993.