A STUDY OF DECISION TREE ALGORITHMS FOR CONTINUOUS ATTRIBUTES

LĒMUMU KOĶA KONSTRUĒŠANAS ALGORITMU IZPĒTE NEPĀRTRAUKTIEM ATRIBŪTIEM

Ieva Boļakova, mag.paed., Daugavpils Pedagogical University,
Parades 1–412, Daugavpils LV 5400, Latvia,Phone: 54 25321, E–mail: ievina@dpu.lv

Abstract. *Nowadays a lot of different algorithms for decision trees construction exist. With the help of these algorithms one can make classification of both discrete and continuous data. The aim of this paper is to explore decision tree algorithms for continuous attributes. There are investigated CART (Breiman et al., 1984) and C4.5 (Quinlan, 1992) in this paper. The comparison of these methods was done in the process of exploration. As a result of the usage of both algorithms, the conclusions about CART and C4.5 utilization advantages were drawn.*

Keywords: *decision trees, CART, C4.5.*

The decision making cannot be regarded as an isolate mechanism or action. It is just one of many stages in the evolution of purposeful activity. We may also state that one stage of some process cannot be considered as more important than all activity as a whole. Before a decision is made, various important activities should be carried out such as data acquisition, representation and classification, and thereafter the formation of corresponding. Thus the outcome of this activity is any kind of choice from some set of alternatives.

It is very convenient to use decision trees for data classification. At present, different decision tree construction algorithms are known. They are being improved successfully and at the same time new more effective methods are being searched for.

Some decision tree algorithms are intended to classify discrete data, but others to classify numeric information. However, there are many situations when a data set consists from both discrete and continuous attributes. For example, the description of a person might include his weight in kilograms, with a value such as 70.5 kg, and the color of eyes whose value may be "brown", "blue" etc. [3] In this case we have to choose a decision tree construction method corresponding to the situation.

In this paper we aim to discuss specific decision tree algorithms for continuous attributes and to give a comparison of them.

Let's choose CART [1] and C4.5 [2] algorithms for discussion. In what follows we will describe them briefly.

The basic outline of CART **(Classification and Regression Trees)**	The basic outline of **C4.5**
1. Found a set of binary questions, where each question is of the form {Is $x \in A$?}, $A \subset X$. This set of binary questions is made for each attribute [1]. 2. A goodness of split criterion $\phi(s,t)$ is then calculated for each of binary questions: $\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$ Suppose that for any node t, there is a candidate split s	A *testT* is chosen, based on a single attribute that has one or more mutually exclusive outcomes $O_1, O_2, \ldots O_n$. T is partitioned into subsets $T_1, T_2, \ldots T_n$, where T_i contains all the cases in set T that have outcome O_i of the chosen test [2]. 1. Consider all tests that divide T into two or more subsets. Score each test according to how well it splits

ISBN 9984 - 585 - 36 - 0

of the node which divides it into t_L and t_R such that a proportion p_L of the cases in t go into t_L and a proportion p_R go into t_R (Figure 1). [1]

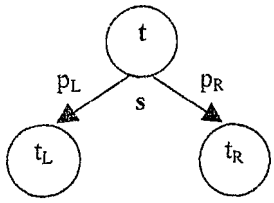

Figure 1

In its turn $i(t)= -\sum_j p(j,t)\log p(j,t)$, where

j is the number of classes.

3. Choose the optimal binary question for each attribute (split criteria is the largest).

4. Find that split s* which gave the largest decrease in impurity – after that manner we find a question which will be in the root node.

5. Repeat the first four steps for each next non–conclusion node.

up the examples. The default test for continuous attributes is A≤t, where A is a continuous attribute, with two outcomes, *true* and *false*. To find the threshold t that maximizes the splitting criterion, the cases in T are sorted by their values of attribute A to give ordered distinct values v_1, v_2, ... v_n. For every pair of adjacent values a potential threshold $t=(v_i+v_{i+1})/2$ is calculated.

2. The threshold that yields the best value of the splitting criterion is then selected.
The default splitting criterion used by this algorithm is gain criterion measured in bits.

$$\text{info(S)}=-\sum_{j=1}^{k}\frac{freq(C_j,S)}{|S|}*\log_2\left(\frac{freq(C_j,S)}{|S|}\right),$$

where info(S) determines the average amount of information needed to identify the class of a case in S [2].

S is any set of cases; $\text{freq}(C_j,S)$ is the number of cases in S that belong to class C_j; |S| is the number of cases in set S.

Gain criteria measures the information that is gained by partitioning T in accordance with the *test X*:

gain(X)=info(T)−info_X(T) [2].

3. Divide the examples into subsets and run this procedure recursively on each subset.

Conclusions. The algorithms were tested in several data sets. The results of both methods sufficiently depend on the choice of appropriate attribute that divides the given data set into subsets. There is calculated some split criterion in both methods so as to make a better choice.

The algorithm C4.5 calculates a split criterion for the threshold value. Thus the in–between values of the attribute are considered. It is not necessary to examine all such thresholds. If all cases with value v_i and with adjacent value v_{i+1} belong to the same class, a threshold between them cannot lead to a partition that has the maximum value of the criterion [3]. This is the reason why the calculus became simpler.

It is necessary to beware of large decision tree construction in both algorithms. For this purpose the following rule should be observed: the splitting process is stopped when further growth of other class impurity decreasing is not possible.

The advantage of both algorithms CART and C4.5 is that they can be employed to construct decision trees for data sets with discrete and continuos information.

References

1. Breiman L., Freidman J., etc. Classification and Regression Trees. – Wadsworth International, Monterey, 1984
2. Quinlan J.R. C4.5: Programs for Machine Learning – The Morgan Kaufmann Series in Machine Learning, Pat Langley, Series Editor, 1992

ISBN 9984 - 585 - 36 - 0

3. Quinlan J.R. Improved Use of Continuos Attributes in C4.5. – Journal of Artifical Intelligence Research 4, 3/96, 77–90, http://www.cs.washington.edu/research/jair/ abstracts/quinlan96a.html

PRODUCTIVITY OF STUDYING PROCESS USING IT
STUDIJU PROCESA PRODUKTIVITĀTE, IZMANTOJOT INFORMĀCIJAS TEHNOLOĢIJAS (IT)

Sarma Cakula, mag.paed. teacher of Vidzeme University College,
Terbatas 10, Valmiera, Latvia, LV 4200, Phone: 371 42 23024, E–mail:
sarma@valmiera.lanet.lv, Fax 371 4223029

Abstract. *It is possible to involve students in learning process more actively using the new information technologies, research method and co–operation. The paper contains theoretical base of student research work as a component of studying process in higher education establishments using IT. The research investigates student personality development and interconnection with productivity of studying process. The author analyses researches on productive interaction in the context of computer–supported collaborative learning in science, computers in the community of classrooms, a sociocultural perspective on the human–technology link and computer–mediated communication. The paper contains empirical research results about productivity of studying process on an experimental base increasing a part of the research work and problem solving using IT and collaboration in studying process of Computer science course in Vidzeme University College.*

Introduction

Computers have become almost ubiquitous over the last years of the twentieth century and one thing that is clear about the twenty–first century. Computers will play an increasingly significant role in our working lives and leisure environments. The question is what the computer has to offer as a technology for supporting education more generally. Information technology (IT) is the study or use of processes (esp. computers etc) for storing, retrieving, and sending information *(Oxford Dic. 1994, 327)*. Many psychologists and educators have a view that IT is the beginning of radical upturn in the education *(Light P., Colbourn C., Light V. 1997)*. But here we can see different tendencies. A great deal of software developed for school use has one way: breaking desired learning goals into small steps and relying on reward, repetition and contingent depending of different levels to impart various skills. It is software developed specifically for individual use *(Howe et.al 1992)*. The next are 'Intelligent Tutoring Systems', which shape a teaching strategy. But these are only small part how to use the IT in the learning process. It is necessary for students in studying process not only to learn special courses, but also acquire skills for professional work, different forms of co–operation and communication. It contains many formal and informal communities, group work in the classroom for special problem solving. There is a way for free education and we need to talk about social dimension in learning process using IT *(Cakula S. 1998)*

Using IT in learning process

It is popular to use the computer as a tool in learning process. One of the most effective tasks in learning process is the research. IT could be used in individual work searching for information, writing papers, using practical programs to develop special knowledge and skills. Recent interest to

ISBN 9984 - 585 - 36 - 0