A Method of Processing Unknown Attribute Values by ID3

Pavel Brazdil\textsuperscript{1} and Ivan Bruha\textsuperscript{2}

\textsuperscript{1} LIACC-CIUP  
4100 Porto, Portugal  
\textit{Email: pbrazdil@ciup1.ncc.up.pt}

\textsuperscript{2} McMaster University  
Hamilton, Ont., Canada, L8S 4K1  
\textit{Email: bruha@mcmaster.ca}

Abstract

The original inductive learning algorithm ID3 assumes that all attribute values of all examples are available both within learning and subsequent classification process. There have been designed and implemented several techniques of processing of unknown attribute values by ID3. This paper firstly briefly surveys these techniques, then discusses the authors' modified technique for unknown value processing, and compares some of the above techniques.

Keywords: inductive learning, decision tree, unknown attribute value

1. Introduction

Although the underlying idea of inducing decision trees by ID3 [3] is simple, powerful, and efficient, the algorithm itself does not behave competently if some examples have one or more attribute values missing. Pre-pruning and post-pruning techniques are useful for processing of noisy data but not for missing attribute values. In order to deal with such real-world problems, ID3 has been extended by numerous techniques of processing of unknown attribute values. Survey of these approaches can be found in [4]. Section 2 of this paper presents important terminology and methodology used by the above techniques.

Notwithstanding [4] presents quite a few techniques for unknown attribute values processing, we have - after reading [2] - decided to explore additional possibilities and modifications of the above techniques. [2] compares symbolic and neural net learning algorithms, namely ID3 and the backpropagation algorithm for multi-layer perceptrons. In case of ID3, [2] uses one of the best unknown value processing technique that was experimentally verified in [4]. Nevertheless, [2] exhibits that backpropagation always performs better than ID3 in case of examples with missing attribute values.

Thus, in order to "help" somehow ID3 in unknown attribute value processing, we have designed and implemented a modified technique that improves the performance of ID3 in such processing. Section 3 discusses our method, section 4 experimental results and comparisons.

2. Survey of techniques

There are three places in ID3 where routines for processing unknown attribute values can be situated:

(a) within the criterion for selecting the most informative attribute;
(b) when partitioning the training set according to the selected attribute;
(c) during classification of a new (unseen) object.

[4] presents the following routines for each situation. Each routine is labelled by a letter, thus combination of these routines yields a technique for processing unknown attribute values, labelled by three letters.

(a) When evaluation a criterion for selecting the most informative attribute (say the well-known gain criterion), the unknown values of an attribute $A$ can be process as follows:

\begin{itemize}
  \item Ignore training examples with unknown value of $A$.
\end{itemize}
3. Our modified technique for processing unknown attribute values by ID3

Although [2] exploits the technique R-F-F in ID3 for processing unknown attribute values, it can be observed that ID3 is losing in comparison with the back-propagation algorithm in all domains that have been applied. In order to improve the performance of ID3 we have taken the technique R-F-F and modified it as follows.

(a) As for the attribute selection, we use Quinlan's routine R. More specifically, the residual information of the attribute A (whose minimum value corresponds to maximum information gain, i.e. the most informative attribute) is calculated as follows:

$$I_{res}(A) = -\sum_{v \neq U} f_v \sum_{r} (K_{vt} + K_{ut} f_v) / K_v \log_2 \left( \frac{K_{vt} + K_{ut} f_v}{K_v} \right)$$

where

$v \neq U$ means the sum over all known values of A,

$f_v$ is the relative frequency of the attribute value $A_v$ calculated above the known values of the attribute A,

$K_{vt}$ is the number of training examples of the class $C_r$ with the attribute A equal to $A_v$,

$K_{ut}$ is the number of examples of the class $C_r$ having an unknown value of A,

$K_v$ is the number of training examples with $A=A_v$:

$$K_v = \sum_r K_v$$

(b) Partitioning a training set has been modified as follows. If there are training examples with unknown value of the selected attribute A, then we generate a separate branch for the unknown value of A (as if it were a regular value of A), and simultaneously add a fraction of each training example with unknown value of A to each training subset based on the frequency $f_v$, i.e. the subset for the attribute value $A_v$ will comprise $K_v + K_{ut} f_v$ examples, where $K_{ut}$ is the number of training examples with missing value of A:

$$K_{ut} = \sum_r K_{ut}$$

One can observe that we more or less combine the routines U and F. The underlying idea for that is we consider and process unknown attribute values as a part of application.
I I

[89x448]factor

[90x470]judge that the known values mostly contribute to this

[91x617]branch (routine

[91x173]25%

[91x183]the filter to get data sets with

[91x544]decision. Hence, the ratio

[91x215]randomly changes attribute values to unknown ones. The

[91x544]gation learning algorithm for neural net.

[91x268]poorly namely on this set in comparison to the backpropa-

[91x278]our experiments, since [2] indicates that ID3 behaved quite

[91x385]To summarize, our modified technique could be

[123x375]as

[124x428]\text{[I]}:\]

[125x447]as one can observe the results of our experiments

[126x459]and those in [2] do not match exactly (evidently due to

[127x447]high randomness in data) but they globally exhibit identic-

[128x459]al characteristics of performance of ID3 with

[129x447]technique: the classification accuracy decreases for a training set

[130x459]with larger percentage of missing attribute values.

[131x480]L)

[131x194]one can observe from Fig. 1, we have run

[136x321]to the probability

[138x406]\text{MIL} - 1)

[144x183]5%,

[146x406]1)

[146x406]-(U/F)-\text{UF})\). We have selected the percentage of unknown

[152x639]unseen pattem (object) has miss-

[152x459]as the

[154x639]its

[156x649]classification in our

[156x659]firstly

[158x670]10.

[161x618]all

[163x618]branches for A are

[165x618]explored. More precisely, a token $t_f$ (which is equal to 1 at

[168x618]the tree root) is distributed across all possible values, i.e.

[169x618]each branch $A_k$ is explored using the token value $t_f$. This

[169x618]exploration and distribution continues until leaves of

[170x618]the tree are reached. The token values at the leaves are

[171x607]modified technique is done

[173x618]are

[174x618]weighted by confidence factors and summed for each

[175x618]class, and the unseen pattern is classified to the class with

[176x607]as that the

[178x607]L)

[181x607]in comparison to the backpropa-

[182x607]validation; more specifically, N has been set up to 10, and

[183x639]the

[186x639]unseen pattem is classified to the
class with the highest value.

The latter possibility is a modification of the

[186x639]routine

[189x607]our technique is about

[190x639]R-(U/F)-(UF)

[191x607]smaller percentage of missing attribute values

[192x639]R-(U/F)-(UF)

[193x607]technique. especially for training

[194x639]technique

[195x607]to

[196x607]resulting from Fig. 1

[197x607]accuracy decreases for a training set

[198x607]with

[199x607]larger percentage of missing attribute values.

In this area, the performance of our technique is about 5% better than that of the original one, i.e. the combined

[200x607]one

[201x607]prime exhibits our

[202x607]modification to Quinlan's original routine.

4. Experiments and results

We have used the well-known soybean data set in

[204x607]our experiments, since [2] indicates that ID3 behaved quite

[206x607]poorly namely on this set in comparison to the backpropagation

[209x607]algorithm for neural net.

In order to emulate missing attribute values, we have run the original soybean data through a filter which randomly changes attribute values to unknown ones. The filter procedure has the percentage of missing values as its parameter. As one can observe from Fig. 1, we have run the filter to get data sets with 5%, 10%, 15%, 20%, and 25% unknown values.

Furthermore, in order to get statistically significant results we have run widely used method of N-cross-validation; more specifically, N has been set up to 10, and the number of runs for each configuration has been also equal to 10.

We have carried out two streams of experiments: firstly the Quinlan's technique $R$-$F$-$F$, and then our modification $R$-$\{U/F\}$-$\{U/F\}$. We have selected the classification accuracy as the only criterion for comparison of these methods. The results are on Fig. 1 which exhibits the classification accuracy vs. percentage of unknown values (from 0 to 25%, emulated by our filter).

Next, we have selected the percentage of unknown values equal to 5% (resulting from Fig. 1 as an optimal magnitude) and carried out a similar set of experiments where we measured the classification accuracy vs. the size of training set, see Fig. 2.

5. Conclusion

As one can observe the results of our experiments and those in [2] do not match exactly (evidently due to high randomness in data) but they globally exhibit identical characteristics of performance of ID3 with $R$-$F$-$F$ technique: the classification accuracy decreases for a training set with larger percentage of missing attribute values.

Our modified technique has better performance than the $R$-$F$-$F$ technique, especially for training sets with smaller percentage of missing attribute values (5 to 10%). In this area, the performance of our technique is about 5% better than that of the original one, i.e. the combined routine $\{U/F\}$ works better for smaller percentage of missing values. In our opinion, generating a separate branch for unknown values has its reason for smaller frequency of missing values, since only a few attributes in a training set have unknown values, i.e. separate branches for unknown values are generated only in some cases. However, if the percentage of unknown attribute values is larger (greater than 10%), then evidently all (or almost all) attributes have unknown value for some training examples, and consequently separate branches for unknown values are generated for (almost) all attributes; thus, only the routine $U$ is exploited during classification.

Fig. 2 does not seem to bring any significant results, since both techniques have almost identical performance for various sizes of training sets, i.e. our techni-
que is more or less 5% better than the original one for all the sizes.

We have found that modification of existing techniques for unknown attribute value processing can improve (at least slightly) the over-all performance of the inductive learning algorithm ID3. However, domain of the training data is also very important since all these techniques depend on statistical character of data.

References