An Application of Approximated Entropy Measures in Decision Tree Induction

Sinh Hoa Nguyen
Polish-Japanese Institute of
Information Technology
Koszykowa 86, 02008, Warszawa, Poland
e-mail: hoa@mimuw.edu.pl

Hung Son Nguyen
Institute of Mathematics
Warsaw University
Banacha 2, Warsaw 02097, Poland
e-mail: son@mimuw.edu.pl

Abstract

The main task in decision tree construction algorithms is to find the "best partition" of set of objects. We consider the problem of searching for optimal binary partition of continuous attribute domain for large data sets stored in relational data bases (RDB). The straightforward approach to optimal partition selection with respect to entropy measure (which evaluates the quality of a partition) needs $O(N)$ simple queries, where N is the number of pre-assumed partitions. We present new approximated entropy measures that allow to construct the partition very close to optimal, using only $O(\log N)$ simple queries.

Keywords: Data mining, decision tree, approximate entropy measure, large databases.

1 Introduction

Searching algorithms for optimal partitions of real value attributes (features) problem, defined by so called cuts, has been studied by many authors (see e.g., [1, 2, 4, 11]). The main goal of such algorithms is to discover *cuts* which can be used to synthesize decision trees or decision rules of high quality with respect to some quality measures (e.g., quality of classification of new unseen objects, quality defined by the decision tree height, support and confidence). In general, all those problems are hard from computational point of view (e.g., it has been shown in [4] that the searching problem for minimal and consistent set of cuts is NP-hard). Hence, numerous heuristics have been developed for approximate solutions of these problems. These heuristics are based on some measures estimating the quality of extracted cuts.

In this paper we consider the entropy based measure which has been used firstly in ID3 methods [11] and many other. We consider a searching problem for optimal partition of real value attributes assuming that the large data table is represented in a relational data base. The critical factor for time complexity of algorithms solving the discussed problem is the number of simple SQL queries of the form

```
SELECT COUNT
FROM aTable
WHERE (anAttribute BETWEEN
        aValue1 AND aValue2)AND
      (some additional conditions)
```

(related to some interval of attribute values) necessary to construct such partitions. We assume the answer time for such queries does not depend on the interval length (this assumption is satisfied in some existing data base servers). Using straightforward approach to optimal partition selection (with respect to a given measure), the number of necessary queries is of order $O(N)$, where N is the number of preassumed parts of the searching space partition.

In previous papers [6, 7, 8] we presented new algorithm for decision tree induction from large data bases and some approximated discernibility measures that allow to construct the partition very close to optimal, using only $O(\log N)$ simple queries. In this paper we will show that one can extend this approach for entropy measure. We present two approximated entropy measures under two different assumption about dependency between cuts.

2 Basic notions

An *information system* [9] is a pair $\mathbb{A} = (U, A)$, where U is a non-empty, finite set called the *universe* and A is a non-empty finite set of *attributes (or features)*, i.e. $a : U \to V_a$ for $a \in A$, where V_a is called *the value set of* a. Elements of U are called *objects or records*. Two objects $x, y \in U$ are said to be discernible by attributes from A if there exists an attribute $a \in A$ such that $a(x) \neq a(y)$. Any information system of the form $\mathbb{A} = (U, A \cup \{dec\})$ is called *decision table* where $dec \notin A$ is called *decision attribute* (or decision for short). Without loss of generality we assume that $V_{dec} = \{1, \ldots, d\}$. Then the set $DEC_k = \{x \in U : dec(x) = k\}$ will be called the k^{th} *decision class of* $\mathbb{A}$ for $1 \leq k \leq d$.

Any pair (a, c), where a is an attribute and c is a real value, is called *a cut* (if the attribute $a \in A$ has been uniquely specified, then cut can be denoted simply by any $c \in \mathbb{R}$. We say that *"the cut (a, c) discerns a pair of objects x, y"* if either $a(x) < c \leq a(y)$ or $a(y) < c \leq a(x)$.

The decision tree for a given decision table is (in simplest case) a binary directed tree with *test functions* (i.e. boolean functions defined on the information vectors of objects) labelled in internal nodes and decision values (i.e. domain of d) labelled in leaves.

In this paper, we consider decision trees using cuts as test functions. Every cut (a, c) is associated with test function $f_{(a,c)}$ such that

$$\forall_{u \in U} f_{(a,c)}(u) = 1 \Leftrightarrow a(u) > c$$

The typical algorithm for decision tree induction can be described as follows:

1. For a given set of objects U, select a cut (a, c_{Best}) of high quality among all possible cuts and all attributes;

2. Induce a partition U_1, U_2 of U by (a, c_{Best}) ;

3. Recursively apply Step 1 to both sets U_1, U_2 of objects until some stopping condition is satisfied.

3 Decision tree construction from Decision tables

Developing some decision tree induction methods (see [2, 11]) and some supervised discretization methods (see [1, 4]), we should often solve the following problem:

Problem *For a given real value attribute a and set of candidate cuts $\{c_1, ..., c_N\}$, find a cut (a, c_i) belonging to the set of optimal cuts with high probability.*

Usually, we use some *measure (or quality functions)* $F : \{c_1, ..., c_N\} \to \mathbb{R}$ to estimate the quality of cuts. For a given measure F, the *straightforward searching algorithm* for the best cut should compute the values of F for all cuts: $F(c_1), .., F(c_N)$. The cut c_{Best} which optimizes (i.e. maximizes or minimizes) the value of function F is selected as the result of searching process.

In next sections we recall the most frequently used measures for decision tree induction like *"Entropy Function" and "Discernibitity Measure"*, respectively. First we fix some notations. Let us consider the attribute a and the set of all relevant cuts $\mathbf{C}_a = \{c_1, ..., c_N\}$ on a.

Definition 1 *The d-tuple of integers $\langle x_1, .., x_d \rangle$ is called class distribution of the set of objects $X \subset U$ iff $x_k = card(X \cap DEC_k)$ for $k \in \{1, ..., d\}$. If the set of objects X is defined by $X = \{u \in U : p \leq a(u) < q\}$ for some $p, q \in \mathbb{R}$ then the class distribution of X can be called* **the class distribution of interval** *$[p; q)$.*

Figure 1: The illustration of entropy(top) and discernibility(bottom) measure

3.1 Entropy methods

This concept uses class-entropy as a criterion to evaluate the list of best cuts which together with the attribute domain induce the desired intervals. The class information entropy of the set of N objects X with class distribution $\langle N_1, ..., N_d\rangle$, where $N_1 + ... + N_d = N$, is defined by

$$Ent(X) = -\sum_{j=1}^{d} \frac{N_j}{N} \log \frac{N_j}{N}$$

Hence, the class information entropy of the partition induced by a cut point c on attribute a is defined by

$$E\left(a, c; U\right) = \frac{|U_L|}{n} Ent\left(U_L\right) + \frac{|U_R|}{n} Ent\left(U_R\right)$$

where $\{U_1, U_2\}$ is a partition of U defined by c. For a given feature a, the cut $c_{\min}$ which minimizes the entropy function over all possible cuts is selected see Figure 1. There is a number of methods based on information entropy theory reported in [3, 11].

3.2 Maximal Discernibility Principle

In Boolean reasoning methods, cuts are treated as Boolean variables and the problem of searching for optimal set of cuts can be characterized by a Boolean function $f_{\mathbb{A}}$ (where $\mathbb{A}$ is a given decision table). It has been shown that the quality of cuts can be measured by their *discernibility properties* (see [4]). Intuitively, the inner energy of the set of objects $X \subset U$ can be defined by the number of pairs of objects from X to be discerned called $conflict(X)$. Let $\langle N_1, ..., N_d\rangle$ be a class distribution of X, then $conflict(X)$ can be computed by

$$conflict(X) = \sum_{i<j} N_i N_j$$

The cut c which divides the set of objects U into U_1, and U_2 is evaluated by $W(c) = conflict(U) - conflict(U_1) - conflict(U_2)$, i.e. the more is number of pairs of objects discerned by the cut (a, c), the larger is chance that c can be chosen to the optimal set of cut. This measure is called *discernibility measure* and the algorithm is called Maximal-Discernibility heuristics or *the MD-heuristics.*

4 Divide and Conquer Algorithm

The main idea is to apply the *"divide and conquer"* strategy to determine the best cut $c_{Best} \in \mathbf{C} = \{c_1, ..., c_n\}$ with respect to a given quality function $F : \mathbf{C} \longrightarrow R^+$.

First we divide the interval containing all possible cuts into k intervals ($k = 2, 3$, etc.). Next we will use some *approximate measures* $\widetilde{F}$ to predict the interval which most probably contains the best cut with respect to the measure F. This process is repeated until the considered interval consists of one cut. Then the best cut can be chosen between all visited cuts. The problem arises how to define the approximate measure $\widetilde{F}$ evaluating the quality of the interval $[c_L, c_R]$ having class distributions: $(L_1, ..., L_d)$ in $(-\infty, c_L)$; $(M_1, ..., M_d)$ in $[c_L, c_R)$; and $(R_1, ..., R_d)$ in $[c_R, \infty)$. This measure should estimate the quality of the best cut among those belonging to the interval $[c_L, c_R]$. The details of our algorithm are described in Figure 2.

One can see that to determine the value $Eval\left([c_L, c_R], \alpha\right)$ we need to have the class

```
ALGORITHM: Searching for semi-optimal cut
PARAMETERS: $k \in \mathbb{N}$ and $\alpha \in [0;1]$. The approximate measure $Eval([c_L; c_R]$ for any
    interval $[c_L, c_R]$
INPUT: attribute $a$; the set of candidate cuts $\mathbf{C}_a = \{c_1, .., c_N\}$ on $a$;
OUTPUT: The optimal cut $c \in \mathbf{C}_a$

begin
  $Left \leftarrow \min$; $Right \leftarrow \max$;
  while ($Left < Right$)
    1.Divide $[Left; Right]$ into $k$ intervals with equal length by $(k+1)$ boundary points
      i.e.,
          $p_i = Left + i * \frac{Right - Left}{k}$;
      for $i = 0, .., k$.
    2.For $i = 1, .., k$ compute approximate measure $Eval([c_{p_{i-1}}; c_{p_i}])$. Let $[p_{j-1}; p_j]$ be the
      interval with maximal value of $Eval(.)$;
    3.$Left \leftarrow p_{j-1}$; $Right \leftarrow p_j$;
  endwhile;
  Return the cut $c_{Left}$;
end
```

Figure 2: The divide and conquer algorithm for optimal cuts

distributions $(L_1, ..., L_d)$, $(M_1, ..., M_d)$ and $(R_1, ..., R_d)$ of the attribute a in $(-\infty, c_L)$, $[c_L, c_R)$ and $[c_R, \infty)$. This requires only $O(d)$ simple SQL queries of the form:

```
SELECT COUNT FROM DecTable
WHERE (attribute_a
   BETWEEN value_1 AND value_2)
   AND (dec = i)
```

Hence the number of queries required for running our algorithm is of order $O(dk \log_k N)$. In practice we set $k = 3$ because the function $f(k) = dk \log_k N$ over positive integers is taking minimum for $k = 3$. For $k > 2$, instead of choosing the best interval $[p_{i-1}, p_i]$, the algorithm can select the best union $[p_{i-m}, p_i]$ of m consecutive intervals in every step for a predefined parameter $m < k$. The modified algorithm needs more – but still $O(\log N)$ simple queries only.

Let us consider an arbitrary cut c lying between c_L and c_R and let us assume that $\langle x_1, x_2, ..., x_d \rangle$ is a class distribution of the interval $[c_L; c]$. We will construct the approximate measure using one of two specific probabilistic models for distribution of objects in the interval $[c_L, c_R]$ described as following:

Independency assumption: $x_1, x_2, ..., x_d$ are independent random variables with uniform distribution over sets $\{0, ..., M_1\}$, ..., $\{0, ..., M_d\}$, respectively

Full dependency assumption: $x_1, ..., x_d$ are proportional to $M_1, ..., M_d$, i.e.

$$\frac{x_1}{M_1} \simeq \frac{x_2}{M_2} \simeq ... \simeq \frac{x_d}{M_d}$$

4.1 Approximated Discernibility Measure

In previous papers [6, 7] we have show that the *Approximated Discernibility Measure* can be defined by

$$\widetilde{W}([c_L; c_R]) = E([c_L; c_R]) + \Delta \qquad (1)$$

where

$$E([c_L; c_R]) = \frac{W(c_L) + W(c_R) + conflict([c_L; c_R])}{2}$$

and

1. under independency assumption,

   $\Delta = \alpha \cdot \sqrt{D^2(W(c))}$

   for some $\alpha \in [0;1]$;

2. under full dependency assumption

   $\Delta = \frac{[W(c_R) - W(c_L)]^2}{8 \cdot conflict([c_L;c_R])}$

In this paper we will extend this result for entropy measure.

5 Approximate Entropy Measures

In previous sections, the discernibility measure has been successfully approximated. The experimental results show that the decision tree or discretization of real value attributes constructed by means of approximate discernibility measures (using small number of SQL queries) are very close to those which are generated by the exact discernibility measure (but using large number of SQL queries). In this section, we would like obtain similar results for entropy measure.

Recall that in the standard Entropy-based methods (see e.g., [11]) we need the following notions:

1. *Information measure* of object set U

$$\begin{aligned} Ent(U) &= -\sum_{j=1}^{d} \frac{N_j}{N} \log \frac{N_j}{N} \\ &= -\sum_{j=1}^{d} \frac{N_j}{N} (\log N_j - \log N) \\ &= \frac{1}{N} \left(h(N) - \sum_{j=1}^{d} h(N_j) \right) \end{aligned}$$

where $h(x) = x \log x$.

2. *Information Gain* $Gain(a,c;U)$ over the set of objects U received by the cut (a,c) is defined by

$$Ent(U) - \left(\frac{|U_L|}{|U|} Ent\left(U_L\right) + \frac{|U_R|}{|U|} Ent\left(U_R\right) \right)$$

where $\{U_L, U_R\}$ is a partition of U defined by c. We have to chose such a cut (a,c) that maximizes the *information gain* $Gain(a,c;U)$ or minimizes the *Entropy induced by this cut*

$$E\left(a,c;U\right) = \frac{|U_L|}{|U|} Ent\left(U_L\right) + \frac{|U_R|}{|U|} Ent\left(U_R\right)$$

Hence $E\left(a,c;U\right)$ can be computed by:

$\frac{1}{N} \left[h(L) - \sum_{j=1}^{d} h(L_j) + h(R) - \sum_{j=1}^{d} h(R_j) \right]$

where $(L_1, ..., L_d)$, $(R_1, ..., R_d)$ are class distribution of U_L and U_R, respectively.

Analogously to the discernibility measure case [6],[7] the main goal is to predict the quality of the best cut (in sense of Entropy measure) among those cuts c from the interval $[c_L, c_R]$, i.e., $E(a,c;U) = \frac{1}{N} f(x_1, ..., x_d)$ where

$$\begin{aligned} f(x_1, ..., x_d) = h(L+x) - \textstyle\sum_{j=1}^{d} h(L_j + x_j) + \\ + h(R+M-x) - \textstyle\sum_{j=1}^{d} h(R_j + M_j - x_j) \end{aligned}$$

5.1 Full dependency assumption case

In this model, the values $x_1, ..., x_j$ can be replaced by

$$x_1 \simeq M_1 \cdot t; \quad x_2 \simeq M_2 \cdot t; \quad ... \quad x_d \simeq M_d \cdot t$$

where $t = \frac{x}{M} \in [0;1]$ (see Section 4). Hence, the task is to find the minimum of the function

$$\begin{aligned} f(t) = h(L+Mt) - \textstyle\sum_{j=1}^{d} h(L_j + M_j t) + \\ h(R+M-Mt) - \textstyle\sum_{j=1}^{d} h(R_j + M_j - M_j t) \end{aligned}$$

where $h(x) = x \log x$ and $h'(x) = \log x + \log e$. One can compute the derivative of $f(t)$:

$$\begin{aligned} f'(t) &= M \log \frac{L+Mt}{R+M-Mt} \\ &\quad - \sum_{j=1}^{d} M_j \log \frac{L_j + M_j t}{R_j + M_j - M_j t} \end{aligned}$$

Theorem 1 *$f'(t)$ is decreasing function.*

proof Let us compute the second derivative of $f(t)$:

$$\begin{aligned} f''(t) &= \frac{M^2}{L+Mt} - \sum_{j=1}^{d} \frac{M_j^2}{L_j + M_j t} + \\ &\quad \frac{M^2}{R+M-Mt} - \sum_{j=1}^{d} \frac{M_j^2}{R_j + M_j - M_j t} \end{aligned}$$

One can show that $f''(t) \leq 0$ for any $t \in (0,1)$. Recall the well known Minski inequality:

$$\sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2 \geq \left(\sum_{i=1}^{n} a_i b_i \right)^2$$

for any real numbers $a_1, ..., a_n, b_1, ..., b_n$. Using this inequality we have:

$$\sum_{j=1}^{d} (L_j + M_j t) \sum_{j=1}^{d} \frac{M_j^2}{L_j + M_j t} \geq \left(\sum_{j=1}^{d} M_j \right)^2$$

The left hand side $= (L + Mt) \sum_{j=1}^{d} \frac{M_j^2}{L_j + M_j t}$, and the right hand side equals M^2. Hence

$$\sum_{j=1}^{d} \frac{M_j^2}{L_j + M_j t} \geq \frac{M^2}{L + Mt}$$

Similarly we can show that

$$\sum_{j=1}^{d} \frac{M_j^2}{R_j + M_j - M_j \cdot t} \geq \frac{M^2}{R + M - M \cdot t}$$

Hence, for any $t \in (0;1)$ we have $f''(t) \leq 0$. This means that $f'(t)$ is decreasing function in the interval $(0,1)$. □

The following example illustrates the properties of $f'(t)$. Let us consider the interval (c_L, c_R) consisting of 600 objects. The class distributions of intervals $(-\infty; c_l)$, (c_L, c_R) and $(c_R; \infty)$ are following:

	Left	Center	Right
$Dec = 1$	$L_1 = 500$	$M_1 = 100$	$R_1 = 1000$
$Dec = 2$	$L_2 = 200$	$M_2 = 400$	$R_2 = 800$
$Dec = 3$	$L_3 = 300$	$M_3 = 100$	$R_3 = 200$
Sum	$L = 1000$	$M = 600$	$R = 2000$

For this data the graph of derivative $f'(t)$ is shown in the Figure 3.

The proved fact can be used to find the value t_0, for which $f'(t_0) = 0$. If such t_0 exists, the function f has maximum at t_0. Hence one can estimate the Entropy measure of the best cut within the interval $[c_L, c_R]$ (under assumption about strong dependencies between classes) as follows:

Figure 3: The function $f'(t)$

- If $f'(1) \geq 0$ then $f'(t) > 0$ for any $t \in (0;1)$, i.e., $f(t)$ is increasing function. Hence c_R is a best cut.
- If $f'(0) \leq 0$ then $f'(t) \leq 0$ for any $t \in (0;1)$, i.e., $f(t)$ is decreasing function. Hence c_L is a best cut.
- If $f'(0) < 0 < f'(1)$ then locate the root t_0 of $f'(t)$ using "Binary Search Strategy". Then the best cut in $[c_L, c_R]$ can be estimated by $\frac{1}{N} f(t_0)$

5.2 Independency assumption case

In the independency model, one can try to compute the expected value of the random variable $f(x_1, ..., x_d)$ using assumption that for $i = 1, ..., d$, x_i are random variables with discrete uniform distribution over interval $[0, M_i]$.

First, we will show some properties of the function $h(x)$. Let x be a random variable with discrete uniform distribution over interval $[0; M]$. If M is sufficiently large integer, the expected value of $h(a + x) = (a + x) \cdot \log_2(a + x)$ can be evaluated by:

$$E(h(a+x)) \simeq \frac{1}{M} \int_{0}^{M} (a+x) \log(a+x) dx$$

We have

$$\begin{aligned} E(h(a+x)) &= \frac{1}{M} \int_{a}^{a+M} x \log x dx \\ &= \frac{1}{M} \left(\frac{x^2 \log x}{2} - \frac{x^2}{4 \ln 2} \right) \Bigg|_{a}^{a+M} \\ &= \frac{(a+M)h(a+M) - ah(a)}{2M} \\ &\quad - \frac{2a+M}{4 \ln 2} \end{aligned}$$

Now one can evaluate the average value $\frac{1}{N}E(f(x_1, ..., x_d))$ of $E(a, c; U)$ by

$$E(h(L + x)) - \sum_{j=1}^{d} E(h(L_j + x_j)) +$$

$$E(h(R+M-x)) - \sum_{j=1}^{d} E(h(R_j + M_j - x_j))$$

Conclusions

We presented the efficient approach for best cut selection from large data bases based on divide and conquer technique. The crucial problem is to define an approximate measure to evaluate the quality of intervals. In general, the problem seems very hard, but under some assumptions, one can construct such measures. In previous papers we constructed approximate discernibility measures under both independency and full dependency assumptions. We extended those results in this paper for entropy measure. We plan to do some experiments to compare the accuracy and efficiency of different methods.

Acknowledgement This paper has been partially supported by Polish State Committee of Research (KBN) grant No 8T11C02519 and the grant of the Wallenberg Foundation. The research of Sinh Hoa Nguyen has been partially supported by Polish-Japanese Institute of Information Technology.

References

[1] Dougherty J., Kohavi R., Sahami M.: Supervised and unsupervised discretization of continuous features. In Proc. of the 12th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA, 1995, pp. 194–202.

[2] Fayyad, U. M., Irani, K.B.: On the handling of continuous-valued attributes in decision tree generation. Machine Learning **8**, 1992, pp. 87–102.

[3] Fayyad, U. M., Irani, K.B.: The attribute selection problem in decision tree generation. In. Proc. of AAAI-92, San Jose, CA, MIT Press, 1992, pp. 104–110.

[4] Nguyen, H. Son: Discretization Methods in Data Mining. In L. Polkowski, A. Skowron (Eds.): *Rough Sets in Knowledge Discovery* **1**, Springer Physica-Verlag, Heidelberg, 1998, pp. 451–482.

[5] Nguyen H.Son, Skowron A.: Boolean reasoning for feature extraction problems. In: Z.W. Raś and A.Skowron (Eds.): Proc. of ISMIS'97, NC, USA, *Foundation of Intelligent Systems* **LNAI 1325**, Springer Verlag, 1997, pp. 117-126.

[6] Nguyen, H. Son: Efficient SQL-Querying Method for Data Mining in Large Data Bases. Proc. of Sixteenth International Joint Conference on Artificial Intelligence, IJCAI-99, Morgan Kaufmann Publishers, Stockholm, Sweden, 1999, pp. 806-811.

[7] Nguyen, H. Son: On Efficient Construction of Decision tree from Large Databases. Proc. of the 2nd International Conference RSCTC'2000. Springer-Verlag, pp. 316-323.

[8] Nguyen, H. S.: On Efficient Handling of Continuous Attributes in Large Data Bases, Fundamenta Informatica **48(1)**, pp. 61-81

[9] Pawlak Z.: *Rough sets: Theoretical aspects of reasoning about data*, Kluwer Dordrecht, 1991.

[10] Polkowski, L., Skowron, A. (Eds.): *Rough Sets in Knowledge Discovery* **Vol. 1,2**, Springer Physica-Verlag, Heidelberg, 1998.

[11] Quinlan, J. R. *C4.5. Programs for machine learning.* Morgan Kaufmann, San Mateo CA, 1993.

[12] Skowron, A., Rauszer, C.: The discernibility matrices and functions in information systems. In. R. Słowiński (ed.). Intelligent Decision Support – Handbook of Applications and Advances of the Rough Sets Theory, Kluwer Academic Publishers, Dordrecht, 1992, pp. 311–362

[13] Komorowski, J., Pawlak, Z., Polkowski, L. and Skowron, A.: Rough sets: A tutorial. In: S.K. Pal and A. Skowron (eds.), Rough - fuzzy hybridization: A new trend in decision making, Springer-Verlag, Singapore, 1999, pp. 3-98.

[14] Ziarko, W.: Rough set as a methodology in Data Mining. In Polkowski, L., Skowron, A. (Eds.): *Rough Sets in Knowledge Discovery* **Vol. 1,2**, Springer Physica-Verlag, Heidelberg, 1998, pp. 554–576.