PAPER *Surveys on Discovery Science*

Expressive Tests for Classification and Regression**

Shinichi Morishita† *and* **Akihiro Nakaya**†, *Nonmembers*

SUMMARY We address the problem of computing various types of expressive tests for decision trees and regression trees. Using expressive tests is promising, because it may improve the prediction accuracy of trees, and it may also provide us some hints on scientific discovery. The drawback is that computing an optimal test could be costly. We present a unified framework to approach this problem, and we revisit the design of efficient algorithms for computing important special cases. We also prove that it is intractable to compute an optimal conjunction or disjunction.

key words: classification, regression, decision trees

1. Introduction

A decision (resp. regression) tree is a rooted binary tree structure for predicting the categorical (numeric) values of the objective attribute. Each internal node has a test on conditional attributes that splits data into two classes. A record is recursively tested at internal nodes and eventually reaches a leaf node. A good decision (resp. regression) tree has the property that almost all the records arriving at every node take a single categorical value (a numeric value close to the average) of the objective attribute with a high probability, and hence the single value (the average) could be a good predictor of the objective attribute.

Making decision trees [9]–[11] and regression trees [2] has been a traditional research topic in the field of machine learning and artificial intelligence. Recently the efficient construction of decision trees and regression trees from large databases has been addressed and well studied among the database community and the KDD community. For details, see the proceedings of recent ACM SIGMOD or SIGKDD conferences. Computing tests at internal nodes is the most time-consuming step of constructing decision trees and regression trees. In the literature, there have been used simple tests that check if the value of an attribute is equal to (or less than) a specific value.

Using more expressive tests is promising in the sense that it may reduce the size of decision or regression trees while it can retain higher prediction accuracy [3], [8]. The drawback however is that the use of expressive tests could be costly. We consider the following three types of expressive tests for partitioning data into two classes; 1) subsets of categorical values for categorical attributes, 2) ranges and regions for numeric attributes, and 3) conjunctions and disjunctions of tests. We present a unified framework for handling those problems. We then reconstruct efficient algorithms for the former two problems, and we prove the intractability of the third problem.

Manuscript received April 30, 1993.

Manuscript revised January 31, 1995.

†The authors are with the Institute of Medical Science, University of Tokyo.

**A Preliminary version of this paper appeared in *Proceedings of the First International Conference on Discovery Science* (Kyushu, December 1998) Springer, Vol. 1532, pages 40-57.

2. Preliminaries

2.1 Relation Scheme, Attribute and Relation

Let $\mathcal{R}$ denote a relation scheme, which is a set of categorical or numeric attributes. The domain of a categorical attribute is a set of unordered distinct values, while the domain of a numeric attribute is real numbers or integers. We select a Boolean or numeric attribute A as special and call it the *objective* attribute. We call the other attributes in $\mathcal{R}$ *conditional attributes*.

Let B be an attribute in relation scheme $\mathcal{R}$. Let t denote a record (tuple) over $\mathcal{R}$, and let $t[B]$ be the value for attribute B. A set of records over $\mathcal{R}$ is called a *relation* over $\mathcal{R}$.

2.2 Tests on Conditional Attributes

We will consider several types of tests for records in a database. Let B denote an attribute, and let v and v_i be values in the domain of B. $B = v$ is a simple test, and t meets $B = v$ if $t[B] = v$.

When B is a categorical attribute, let $\{v_1, \ldots, v_k\}$ be a subset of values in the domain of B. Then, $B \in \{v_1, \ldots, v_k\}$ is a test, and t satisfies this test if $t[B]$ is equal to one value in $\{v_1, \ldots, v_k\}$. We will call a test of the form $B \in \{v_1, \ldots, v_k\}$ a *test with a subset of categorical values*.

When B is a numeric attribute, $B = v$, $B \leqq v$, $B \geqq v$, and $v_1 \leqq B \leqq v_2 (B \in [v_1, v_2])$ are tests, and a record t meets them respectively if $t[B] = v$, $t[B] \leqq v$, $t[B] \geqq v$, and $v_1 \leqq t[B] \leqq v_2$. We will call a test of the form $B \in [v_1, v_2]$ a *test with a range*.

The negation of a test T is denoted by $\neg T$. A record t meets $\neg T$ if t does not satisfy T. The negation

of $\neg T$ is T.

A conjunction (a disjunction, respectively) of tests $T_1, T_2, \ldots, T_k$ is of the form $T_1 \wedge T_2 \wedge \ldots \wedge T_k$ ($T_1 \vee T_2 \vee \ldots \vee T_k$). A record t meets a conjunction (respectively, a disjunction) of tests, if t satisfies all the tests (some of the tests).

2.3 Splitting Criteria for Boolean Objective Attribute

Splitting Relation in Two

Let R be a set of records over $\mathcal{R}$, and let $|R|$ denote the number of records in R. Let $Test$ be a test on conditional attributes. Let R_1 be the set of records that meet $Test$, while let R_2 denote $R - R_1$. In this way, we can use $Test$ to divide R into R_1 and R_2. Suppose that the objective attribute A is Boolean. We call a record whose A's value is true a *positive* record with respect to the objective attribute A. Let R^t denote the set of positive records in R. On the other hand, we call a record whose A's value is false a *negative* record, and let R^f denote the set of negative records in R. The following diagram illustrates how R is partitioned.

$$\begin{array}{ccc} & R = R^t \cup R^f & \\ \swarrow & & \searrow \\ R_1 = R_1^t \cup R_1^f & & R_2 = R_2^t \cup R_2^f \end{array}$$

The splitting by $Test$ is effective for characterizing the objective Boolean attribute A if the probability of positive records changes dramatically after the division of R into R_1 and R_2; for instance, $|R^t|/|R| \ll |R_1^t|/|R_1|$, and $|R^t|/|R| \gg |R_2^t|/|R_2|$. On the other hand, the splitting by $Test$ is most ineffective if the probability of positive records does not change at all; that is, $|R^t|/|R| = |R_1^t|/|R_1| = |R_2^t|/|R_2|$.

Measuring the Effectiveness of Splitting

It is helpful to have a way of measuring the effectiveness of the splitting by a condition. To define the measure, we need to consider $|R|$, $|R^t|$, $|R^f|$, $|R_1|$, $|R_1^t|$, $|R_1^f|$, $|R_2|$, $|R_2^t|$ and $|R_2^f|$ as parameters, which satisfy the following equations:

$$\begin{array}{ll} |R| = |R^t| + |R^f| & |R| = |R_1| + |R_2| \\ |R_1| = |R_1^t| + |R_1^f| & |R_2| = |R_2^t| + |R_2^f| \\ |R^t| = |R_1^t| + |R_2^t| & |R^f| = |R_1^f| + |R_2^f| \end{array}$$

Since R is given and fixed, we can assume that $|R|$, $|R^t|$, and $|R^f|$ are constants. Let n and m denote $|R|$ and $|R^t|$ respectively, then $|R^f| = n - m$. Furthermore, if we give the values of $|R_1|$ and $|R_1^t|$, for instance, the values of all the other variables are determined. Let x and y denote $|R_1|$ and $|R_1^t|$ respectively. Let $\phi(x, y)$ denote the measurement of the effectiveness of the splitting by condition $Test$. We now discuss some requirements that $\phi(x, y)$ is expected to have.

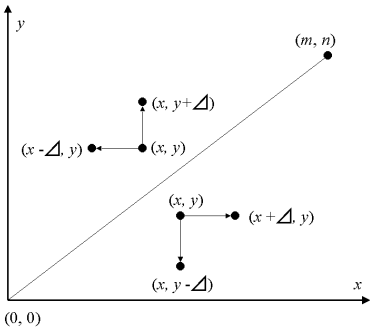

Fig. 1 (x, y), $(x, y + \Delta)$, $(x - \Delta, y)$, $(x, y - \Delta)$, and $(x + \Delta, y)$

We first assume that lower value of $\phi(x, y)$ indicates higher effectiveness of the splitting. It does not matter if we select the reverse order. The splitting by $Test$ is most ineffective when $|R^t|/|R| = m/n = |R_1^t|/|R_1| = y/x = |R_2^t|/|R_2|$, and hence $\phi(x, y)$ should be maximum when $y/x = m/n$.

Suppose that the probability of positive records in R_1, y/x, is greater than that of positive records in R, m/n. Also suppose that if we divide R by another new test, the number of positive records in R_1 increases by Δ ($0 < \Delta \leqq x - y$), while $|R_1|$ is the same. Then, the probability of positive records in R_1, $(y + \Delta)/x$, becomes to be greater than y/x, and hence we want to claim that the splitting by the new test is more effective. Thus we expect $\phi(x, y + \Delta) \leqq \phi(x, y)$. Similarly, since $y/x \leqq y/(x - \Delta)$ for $0 \leqq \Delta < x - y$ we also expect $\phi(x - \Delta, y) \leqq \phi(x, y)$. Figure 1 illustrates points (x, y), $(x, y + \Delta)$, and $(x - \Delta, y)$.

If the probability of positive records in R_1, y/x, is less than the average m/n, then (x, y) is in the lower side of the line connecting the origin and (m, n). See Figure 1. In this case observe that the probability of positive records in R_2, which is $(m - y)/(n - x)$, is greater than m/n. Suppose that the number of positive records in R_1 according to the new test decreases by Δ ($0 < \Delta \leqq x - y$), while $|R_1|$ is unchanged. Then, the number of positive records in R_2 increases by Δ while $|R_2|$ is the same. Thus the splitting by the new test is more effective, and we expect $\phi(x, y - \Delta) \leqq \phi(x, y)$. Similarly we also want to require $\phi(x + \Delta, y) \leqq \phi(x, y)$.

In summary ϕ is expected to satisfy that if $y/x > m/n$, then $\phi(x, y+\delta) \leqq \phi(x, y)$ and $\phi(x-\delta, y) \leqq \phi(x, y)$, oherwise, $\phi(x, y-\delta) \leqq \phi(x, y)$ and $\phi(x+\delta, y) \leqq \phi(x, y)$.

Entropy of Splitting

We present an instance of $\phi(x, y)$ that meets all the requirements discussed so far. Let $ent(p) = -p \ln p - (1 - p) \ln(1 - p)$, where p means the probability of positive records in a set of records, while $(1 - p)$ implies the probability of negative records. Define the *entropy*

$Ent(x, y)$ of the splitting by $Test$ as follows:

$$\frac{x}{n} ent(\frac{y}{x}) + \frac{n-x}{n} ent(\frac{m-y}{n-x}),$$

where $\frac{y}{x}$ ($\frac{m-y}{n-x}$, respectively) is the probability of positive records in R_1 (R_2). This function is known as Quinlan's entropy heuristic [9], and it has been traditionally used as a criteria for evaluating the effectiveness of the division of a set of records. $Ent(x, y)$ is an instance of $\phi(x, y)$. We use the following theorem to show that $Ent(x, y)$ satisfies all the requirements on $\phi(x, y)$.

Theorem 2.1: $Ent(x, y)$ is a concave function for $x \geqq y \geqq 0$; that is, for any (x_1, y_1) and (x_2, y_2) in $\{(x, y) \mid x \geqq y \geqq 0\}$ and any $0 \leqq \lambda \leqq 1$,

$$\begin{aligned} &\lambda Ent(x_1, y_1) + (1-\lambda) Ent(x_2, y_2) \\ \leqq\ & Ent(\lambda(x_1, y_1) + (1-\lambda)(x_2, y_2)). \end{aligned}$$

$Ent(x, y)$ is maximum when $y/x = m/n$.

Proof: See Appendix. ∎

We immediately obtain the following corollary.

Corollary 2.1: Let (x_3, y_3) be an arbitrary dividing point of (x_1, y_1) and (x_2, y_2) in $\{(x, y) \mid x \geqq y \geqq 0\}$; namely, (x, y) is a point on the line segment connecting the two points. Then, $\min(Ent(x_1, y_1), Ent(x_2, y_2)) \leqq Ent(x_3, y_3)$.

For any (x, y) such that $y/x > m/n$ and any $0 < \Delta \leqq x - y$, from the above corollary we have

$$\min(Ent(x, y + \Delta), Ent(x, (m/n)x)) \leqq Ent(x, y),$$

because (x, y) is a dividing point of $(x, y + \Delta)$ and $(x, (m/n)x)$. Since $Ent(x, (m/n)x)$ is maximum,

$$Ent(x, y + \Delta) \leqq Ent(x, y),$$

and hence $Ent(x, y)$ satisfies the requirement $\phi(x, y + \Delta) \leqq \phi(x, y)$. In the same way we can show that $Ent(x, y)$ meets all the requirements on $\phi(x, y)$.

2.4 Splitting Criteria for Numeric Objective Attribute

Consider the case when the objective attribute A is numeric. Let $\mu(R)$ denote the average of A's values in relation R; that is, $\mu(R) = \sum_{t \in R} t[A]/|R|$. Let R_1 denote again the set of records that meet a test on conditional attributes, while let R_2 denote $R - R_1$.

In order to characterize A, it is useful to find a test such that $\mu(R_1)$ is considerably higher than $\mu(R)$ while $\mu(R_2)$ is substantially lower than $\mu(R)$ simultaneously. To realize this criteria, we use the *interclass variance* of the splitting by the test:

$$|R_1|(\mu(R_1) - \mu(R))^2 + |R_2|(\mu(R_2) - \mu(R))^2.$$

A test is more interesting if the interclass variance of the splitting by the test is larger. We also expect that the variance of A's values in R_1 (resp., R_2) should be small, which lets us approximate A's values in R_1 (R_2) at $\mu(R_1)$ ($\mu(R_2)$). To measure this property, we employ the *intraclass variance* of the splitting by the test:

$$\frac{\sum_{t \in R_1} (t[A] - \mu(R_1))^2 + \sum_{t \in R_2} (t[A] - \mu(R_2))^2}{|R|}.$$

We are interested in a test that maximizes the interclass variance and also minimizes the intraclass variance at the same time. Actually the maximization of the interclass variance coincides with the minimization of the intraclass variance.

Theorem 2.2: Given a set of tests on conditional attributes, the test that maximizes the interclass variance also minimizes the intraclass variance.

Proof: See Appendix. ∎

In what follows, we will focus on the maximization of the interclass variance. When R is given and fixed, $|R| (= |R_1| + |R_2|)$ and $\sum_{t \in R} t[A]$ can be regarded as constants, and let n and m denote $|R|$ and $\sum_{t \in R} t[A]$ respectively. If we denote $|R_1|$ and $\sum_{t \in R_1} t[A]$ by x and y, the interclass variance is determined by x and y as follows:

$$x(\frac{y}{x} - \frac{m}{n})^2 + (n - x)(\frac{m-y}{n-x} - \frac{m}{n})^2,$$

which will be denoted by $Var(x, y)$. We then have the following property of $Var(x, y)$, which is similar to Theorem 2.1 for the entropy function.

Theorem 2.3: $Var(x, y)$ is a convex function for $0 < x < n$; that is, for any (x_1, y_1) and (x_2, y_2) such that $n > x_1, x_2 > 0$ and any $0 \leqq \lambda \leqq 1$,

$$\begin{aligned} &\lambda Var(x_1, y_1) + (1-\lambda) Var(x_2, y_2) \\ \geqq\ & Var(\lambda(x_1, y_1) + (1-\lambda)(x_2, y_2)). \end{aligned}$$

$Var(x, y)$ is minimum when $y/x = m/n$.

Proof: See Appendix. ∎

Corollary 2.2: If (x_3, y_3) be an arbitrary dividing point of (x_1, y_1) and (x_2, y_2) such that $n > x_1, x_2 > 0$, then $\max(Var(x_1, y_1), Var(x_2, y_2)) \geqq Var(x_3, y_3)$.

Since the interclass variance has the property similar to the entropy function, in the following sections, we will present how to compute the optimal test that minimizes the entropy, but all arguments directly carry over to the case of finding the test maximizing the interclass variance.

2.5 Positive Tests and Negative Tests

Let R be a given relation, and let R_1 be the set of records in R that meet a given test. If the objective attribute A is Boolean, we treat "true" and "false" as numbers "1" and "0" respectively. We call the test *positive* if the average of A's values in R_1 is greater than or equal to the average of A's values in R; that is, $(\sum_{t \in R_1} t[A])/|R_1| \geqq (\sum_{t \in R} t[A])/|R|$. Otherwise the

test is called *negative*. Thus, when A is Boolean, the probability of positive records in R_1 is greater than or equal to the probability of positive records in R.

The test that minimizes the entropy could be either positive or negative. In what follows, we will focus on computing the positive test that minimizes the entropy of the splitting by the positive test among all the positive tests. This is because the algorithm for computing the optimal positive test can be used to calculate the optimal negative test by exchanging "true" and "false" of the objective Boolean attribute value (or reversing the order of the objective numeric attribute value) in each record.

3. Computing Optimal Tests with Subsets of Categorical Values

Let C be a conditional categorical attribute, and let $\{c_1, c_2, \ldots, c_k\}$ be the domain of C. Among all the positive tests of the form $C \in S$ where S is a subset of $\{c_1, c_2, \ldots, c_k\}$, we want to compute the positive test that minimizes the entropy of the splitting. A naive solution would consider all the possible subsets of $\{c_1, c_2, \ldots, c_k\}$ and select the one that minimizes the entropy. Instead of investigating all 2^k subsets, there is an efficient way of checking only k subsets.

We first treat "true" and "false" as real numbers "1" and "0" respectively. Assume that $\{t \mid t[C] = c_i\}$ is non-empty for simplicity. Otherwise, remove c_i from the domain of C. For each c_i, let μ_i denote the average of A's values of all the records whose C's values are c_i; that is,

$$\mu_i = \frac{\sum_{t[C]=c_i} t[A]}{|\{t \mid t[C] = c_i\}|}.$$

Without loss of generality we can assume that $\mu_1 \geqq \mu_2 \geqq \ldots \geqq \mu_k$, otherwise we rename the categorical values appropriately to meet the above property. We then have the following theorem.

Theorem 3.1: Among all the positive tests with subsets of categorical values, there exists a positive test of the form $C \in \{c_i \mid 1 \leqq i \leqq j\}$ that minimizes the entropy of the splitting.

This theorem is due to Breiman et al.[2]. Thanks to this theorem, we only need to consider k tests of the form $C \in \{c_i \mid 1 \leqq i \leqq j\}$ to find the optimal test. We now prove the theorem by using techniques introduced in the previous section.

Proof of Theorem 3.1

We will prove the case of the minimization of the entropy. The case of maximization of the interclass variance can be shown similarly. With each subset W of $\{c_1, c_2, \ldots, c_k\}$, we associate

$$p(W) = (\ |\{t \mid t[C] \in W\}|, \sum_{t[C]\in W} t[A]\ )$$

in the Euclidean plane. $Ent(p(W))$ is the entropy of the splitting by the test $C \in W$. Consider the set of all points associated with all subsets of $\{c_1, c_2, \ldots, c_k\}$. It is well known that a concave function is minimized at the boundary of the convex hull of all those points. Since we focus on positive tests with subsets of categorical values, the subset W^* minimizing $Ent(p(W)$ can be found by computing the point $p(W)$ that maximizes

$$\sum_{t[C]\in W} t[A] - \lambda|\{t \mid t[C] \in W\}|,$$

where λ is a positive parameter. Consider the equality:

$$\begin{aligned}
&\sum_{t[C]\in W} t[A] - \lambda|\{t \mid t[C] \in W\}| \\
&= \sum_{c_i \in W} (\sum_{t[C]=c_i} t[A] - \lambda|\{t \mid t[C] = c_i\}|) \\
&= \sum_{c_i \in W} |\{t \mid t[C] = c_i\}|(\mu_i - \lambda).
\end{aligned}$$

For the purpose of maximization, we need to exclude from W such c_i that $\mu_i - \lambda < 0$. Thus, $W^* = \{c_i \mid \mu_i \geqq \lambda\}$. Since $\mu_1 \geqq \mu_2 \geqq \ldots \geqq \mu_k$, $W^* = \{c_i \mid 1 \leqq i \leqq j\}$ for some j. ∎

4. Computing Optimal Tests with Ranges or Regions

Let B be a conditional attribute that is numeric, and let I be a range of the domain of B. We are interested in finding a test of the form $B \in I$ that minimizes the entropy (or maximizes the interclass variance) of the splitting by the test. When the domain of B is real numbers, the number of candidates could be infinite. One way to cope with this problem is that we discretize this problem by dividing the domain of B into disjoint sub-ranges, say $I_1, \ldots, I_N$, so that the union $I_1 \cup \ldots \cup I_N$ is the domain of B. The division of the domain, for instance, can be done by distributing the values of B in the given set of records into equal-sized sub-ranges. We then concatenate some successive sub-ranges, say $I_i, I_{i+1}, \ldots, I_j$, to create a range $I_i \cup I_{i+1} \cup \ldots \cup I_j$ that optimizes the criteria of interest.

It is natural to consider the two-dimensional version. Let B and C be numeric conditional attributes. We also simplify this problem by dividing the domain of B (resp. C) into N_B (N_C) equal-sized sub-ranges. We assume that $N_B = N_C = N$ without loss of generality as regards our algorithms. We then divide the Euclidean plane associated with B and C into $N \times N$ pixels. A *grid* region is a set of pixels, and let R be an instance. A record t satisfies test $(B, C) \in R$ if $(t[B], t[C])$ belongs to R. We can consider various types of grid regions for the purpose of splitting a relation in two. In the literature two classes of regions have been well studied [3], [4], [8], [12]. An *x-monotone* region is

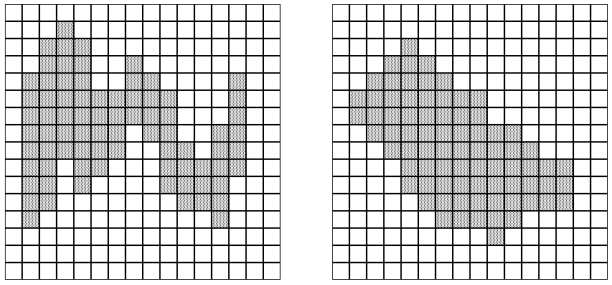

Fig. 2 An x-monotone region (left) and a rectilinear convex region (right)

a connected grid region whose intersection with any vertical line is undivided. A *rectilinear convex* region is an x-monotone region whose intersection with any horizontal line is also undivided. Figure 2 shows an x-monotone region in the left and a rectilinear convex region in the right.

In the case of computing the optimal range by concatenating some consecutive sub-ranges of N sub-ranges, we may consider $O(N^2)$ sequences of successive sub-ranges, but to this end, Katoh [6] presents an $O(N \log N)$-time algorithm.

On the other hand, the number of x-monotone regions and the number of rectilinear convex regions is more than 2^N. It is non-trivial to efficiently find such a region R that minimizes the entropy (maximizes the interclass variance) of the splitting by the test $(B, C) \in R$. Here we review some techniques for this purpose.

Convex Hull of Stamp Points

Let $\mathcal{R}$ denote the family of x-monotone regions or the family of rectilinear convex regions. Let A be the objective attribute. When A is Boolean, we treat "true" and "false" as real numbers "1" and "0". With each region R in $\mathcal{R}$, we associate a *stamp point* (x, y) where $x = |\{t \mid t \text{ meets } (B, C) \in R\}|$ and $y = \sum_{\{t|t \text{ meets } (B,C)\in R\}} t[A]$. Since the number of regions in $\mathcal{R}$ is more than 2^N, we cannot afford to calculate all the point associated, and hence we simply assume their existence.

Let $\mathcal{S}$ denote the set of stamp points for a family of regions $\mathcal{R}$. A *convex polygon* of $\mathcal{S}$ has the property that any line connecting arbitrary two points of $\mathcal{S}$ must itself lies entirely inside the polygon. The *convex hull* of $\mathcal{S}$ is the smallest convex polygon of $\mathcal{S}$. The left in Figure 3 illustrates the convex hull. The upper (lower) half of a convex hull is called the *upper* (*lower*) hull, in short.

Proposition 4.1: Let $R \in \mathcal{R}$ be the region such that test $(B, C) \in R$ minimizes the entropy (or maximizes the interclass variance). The stamp point associated with R must be on the convex hull of $\mathcal{S}$.

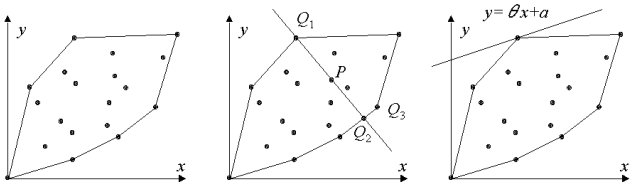

Fig. 3 The left figure presents the convex hull of stamp points. The middle illustrates P, Q_1, Q_2 and Q_3 in Proposition 4.1. The right shows the hand probing technique.

Proof: Otherwise there exists such a point P inside the convex hull of $\mathcal{S}$ that minimizes the entropy. Select any point Q_1 on the convex hull, draw the line connecting P and Q_1, and let Q_2 be another point where the line between P and Q_1 crosses the convex hull. From the concavity of the entropy function, $\min(Ent(Q_1), Ent(Q_2)) \leqq Ent(P)$, and there exists a point Q_3 on the convex hull such that $Ent(Q_3) \leqq Ent(Q_2)$ (see Figure 3). Thus, $Ent(Q_3) \leqq Ent(P)$, which is a contradiction. ∎

If T is the positive (negative, resp.) test that minimizes the entropy among all the positive tests of the form $(B, C) \in R$, from Proposition 4.1 the stamp point associated with T must be on the upper (lower) hull. We then present how to scan the upper hull to search the stamp point that minimizes the entropy.

Hand-Probing

To this end it is useful to use the "hand-probing" technique that was invented by Asano, Chen, Katoh and Tokuyama [1] for image segmentation and was later modified by Fukuda, Morimoto, Morishita and Tokuyama [4] for extraction of the optimal x-monotone region.

For each stamp point on the upper hull, there exists a tangent line to the upper hull at the point. Let θ denote the slope of the tangent line. The right picture in Figure 3 shows the tangent line. Note that the stamp point maximizes $y - \theta x$ among all the stamp points, and let R denote the region that corresponds to the stamp point. We now present a roadmap of how to construct R.

Let $p_{i,j} (1 \leqq i, j \leqq N)$ denote the (i, j)-th pixel in $N \times N$ pixels. A grid region is a union of pixels. Let $u_{i,j}$ be the number of records that meet $(B, C) \in p_{i,j}$, and let $v_{i,j}$ be the sum of the objective attribute values of all the records that satisfy $(B, C) \in p_{i,j}$, which is $\sum_{t \text{ meets } (B,C)\in p(i,j)} t[A]$. Using those notations, we can represent the stamp point associated with R by $(\sum_{p_{i,j} \subseteqq R} u_{i,j}, \sum_{p_{i,j} \subseteqq R} v_{i,j})$, which maximizes $y - \theta x$. Since

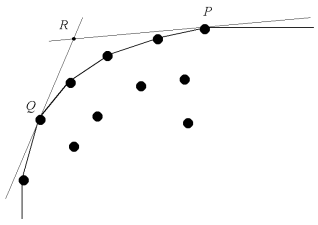

Fig. 4 Guided Branch-and-Bound Search

$$\sum_{p_{i,j} \subseteq\!\!\!= R} v_{i,j} - \theta \sum_{p_{i,j} \subseteq\!\!\!= R} u_{i,j} = \sum_{p_{i,j} \subseteq\!\!\!= R} (v_{i,j} - \theta u_{i,j}),$$

R maximizes $\sum_{p_{i,j} \subseteq\!\!\!= R}(v_{i,j} - \theta u_{i,j})$.

We call $v_{i,j} - \theta u_{i,j}$ the *gain* of the pixel $p_{i,j}$. The problem of computing the region that maximizes the sum of gains of pixels in the region has been studied. For a family of x-monotone regions, Fukuda, Morimoto, Morishita, and Tokuyama presents an $O(N^2)$-time algorithm [4]. For a family of rectilinear convex regions, Yoda, Fukuda, Morimoto, Morishita, and Tokuyama gives an $O(N^3)$-time algorithm [12]. Due to the space limitation, we do not introduce those algorithms. Those algorithms use the idea of dynamic programming, and they connect an interval in each column from lower index i to higher one to generate an x-monotone (or, rectilinear) region.

Since we have an efficient algorithm for generating the region associated with the stamp point on the convex hull at which the line with a slope θ touches, it remains to answer how many trials of hand-probing procedure are necessary to find the region that minimizes the entropy. If n is the number of given records, there could be at most n stamp points on the upper hull, and therefore we may have to do n trials of hand-probing by using n distinct slopes. Next we present a technique that is expected to reduce the number of trials to be $O(\log n)$ in practice.

Guided Branch-and-Bound Search

Using a tangent line with the slope $\theta = 0$, we can touch the rightmost point on the convex hull. Let a be an aribitrary large real number such that we can touch the leftmost point on the convex hull by using the tangent line with slope a. Thus using slopes in $[0, a]$, we can scan all the points on the upper hull. We then perform the binary search on $[0, a]$ to scan the convex hull.

During the process we may dramatically reduce the search space. Figure 4 shows the case when we use two tangent lines to touch two points P and Q on the convex hull, and R denotes the point of intersection of the two lines. Let X be an arbitrary point inside the triangle PQR. From the concavity of the entropy function, we immediately obtain

$$\min\{Ent(P), Ent(Q), Ent(R)\} \leqq Ent(X).$$

If $\min\{Ent(P), Ent(Q)\} \leqq Ent(R)$, we have $\min\{Ent(P), Ent(Q)\} \leqq Ent(X)$, which implies that it is useless to check whether or not there exists a point between P and Q on the hull whose entropy is less than $\min\{Ent(P), Ent(Q)\}$. In practice, most of subintervals of slopes are expected to be pruned away during the binary search. This guided branch-and-bound search strategy has been experimentally evaluated [3], [8]. According to experimental tests the number of trials of hand-probing procedure is $O(\log n)$.

5. Computing Optimal Conjunctions and Disjunctions

Suppose that we are given a set S of tests on conditional attributes. We also assume that S contains the negation of an arbitrary test in S. We call a conjunction *positive* (*negative*, resp.) if it is a positive (negative) test. We will show that it is NP-hard to compute the positive conjunction (the positive disjunction, resp.) that minimizes the entropy among all positive conjunctions (positive disjunctions) of tests in S. Also, it is NP-hard to compute the positive conjunction (positive disjunction) that maximizes the interclass variance.

Let $T_1 \wedge \ldots \wedge T_k$ be a positive conjunction of tests in S. Observe that the entropy (the interclass variance, resp.) of the splitting by $T_1 \wedge \ldots \wedge T_k$ is equal to the entropy (the interclass variance) of the splitting by $\neg(T_1 \wedge \ldots \wedge T_k)$. $\neg(T_1 \wedge \ldots \wedge T_k)$ is equivalent to $\neg T_1 \vee \ldots \vee \neg T_k$, which is a negative disjunction of tests in S. Thus the negation of the optimal positive conjunction gives the optimal negative disjunction. As remarked in Subsection 2.5, computing a negative test can be done by using a way of computing a positive test, and therefore we will prove the intractability of computing the optimal disjunction.

Theorem 5.1: Given a set S of tests on conditional attributes such that S contains the negation of any test in S, it is NP-hard to compute the positive disjunction of tests in S that minimizes the entropy value among all positive disjunctions. It is also NP-hard to compute the positive disjunction that maximizes the interclass variance.

Proof: Here we present a proof for the case of the entropy. The case of the interclass variance can be proved in a similar manner. We reduce the difficulty of the problem to the NP-hardness of MINIMUM COVER [5]. Let V be a finite set, and let C be a collection of subsets of V. A sub-collection $C'(\subset C)$ is a cover of V if any element in V belongs to one of C'. Suppose that C_{min} is a cover that minimizes the number of subsets in it. It is NP-hard to compute C_{min}.

Suppose that V contains a elements, and C contains c subsets of V. We call elements in V *black*. Let b be a number greater than a and c, generate a set W

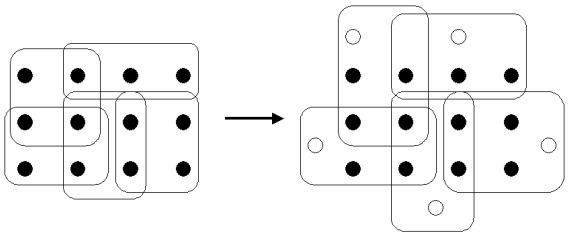

Fig. 5 Each subset is extended with a unique white element.

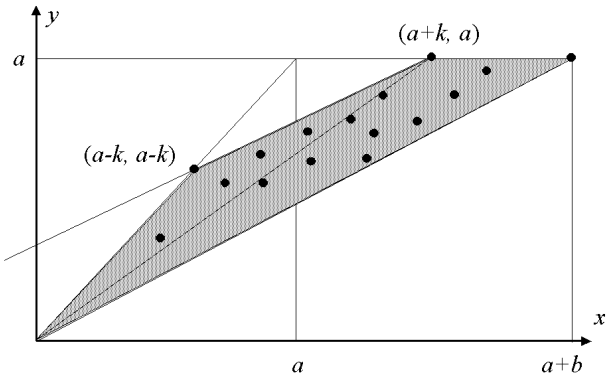

Fig. 6 Points Associated with Sub-collections

of new b elements, and call them *white*. We then extend each subset in C by adding a unique white element that does not appear elsewhere. We do not use $(b-c)$ white elements for this extension. Figure 5 illustrates this operation. In the figure each hyperedge shows a subset in C. After this extension, C and C_{min} become collections of subsets of $V \cup W$.

In what follows, we treat elements in $V \cup W$ as records in a database. We assume that the objective attribute is true (false, resp.) for black (white) records in $V \cup W$. We then identify each subset in C with a test such that all elements in the subset meet the test, while none of elements outside the subset satisfy the test. We also identify a collection $C'(\subseteqq C)$ with the disjunction of tests that correspond to subsets in C'. We then show that the disjunction corresponding to C_{min} minimizes the entropy, which means that finding the optimum disjunction is NP-hard.

With each sub-collection $C'(\subseteqq C)$ such that the disjunction identified with C' is positive, we associate a point (x, y) in an Euclidean plane such that x is the number of records in C', and y is the number of black records in C'. See Figure 6. $Ent(x, y)$ gives the entropy of the disjunction identified with C'. Let k denote the number of subsets in the minimum cover C_{min}. $(a+k, a)$ is associated with C_{min}. We prove that all the points associated with collections of subsets of C fall in the gray region in Figure 6.

All the points lie in the upper side or on the line connecting the origin and $(a+b, a)$, because each point corresponds to a positive disjunction. We show that all the points lie under or on the line between $(a+k, a)$ and $(a-k, a-k)$. To this end, it is enough to prove that any $C' \subset C$ that contains $a-l (l \geqq 1)$ black records must also have at least $k-l$ white records. The proof is an induction on l, and consider the case when $l = 1$. Suppose that the number of white records is less than $k-1$. We can immediately construct a cover of V by adding to C' a subset X that contains the remaining black records. Note that the number of white records in $C' \cup \{X\}$ is less than k, which contradicts the choice of C_{min}. The argument carries over to the case when $l > 1$.

We then prove that $Ent(a-k, a-k) > Ent(a+k, a)$ for $k \geqq 1$.

$$
\begin{aligned}
& Ent(a-k, a-k) \\
&= \frac{a-k}{a+b} ent(\frac{a-k}{a-k}) + \frac{b+k}{a+b} ent(\frac{k}{b+k}) \\
&= \frac{1}{a+b}(-k \ln \frac{k}{b+k} - b \ln \frac{b}{b+k}) \\
&\qquad \text{Because } ent(1) = 0 \\
& Ent(a+k, a) \\
&= \frac{a+k}{a+b} ent(\frac{a}{a+k}) + \frac{b-k}{a+b} ent(\frac{a-a}{b-k}) \\
&= \frac{1}{a+b}(-k \ln \frac{k}{a+k} - a \ln \frac{a}{a+k}) \\
&\qquad \text{Because } ent(0) = 0
\end{aligned}
$$

Let $f(x)$ denote $-k \ln \frac{k}{x+k} - x \ln \frac{x}{x+k}$. We then have $Ent(a-k, a-k) = \frac{1}{a+b} f(b)$ and $Ent(a+k, a) = \frac{1}{a+b} f(a)$. Since $f'(x) = \ln \frac{x+k}{x} > 0$ for $x > 0$. Because $b > a > 0$, we have $f(b) > f(a)$, and hence $Ent(a-k, a-k) > Ent(a+k, a)$.

From Theorem 2.1, $Ent(x, y)$ is maximum at any point (x, y) on the line between $(0, 0)$ and $(a+b, a)$, and $Ent(x, y)$ is a concave function on the gray quadrilateral in Figure 6. Since $Ent(a-k, a-k) > Ent(a+k, a)$, the entropy of any point in the gray quadrilateral is no less than the entropy of $(a+k, a)$. Recall that $(a+k, a)$ corresponds to the positive disjunction associated with C_{min}. Consequently the positive disjunction that minimizes the entropy corresponds to C_{min}. ∎

Acknowledgements

The first author thanks Professor Katoh of Kyoto University for stimulus discussion on the topic of this paper. He is also indebted to the anonymous referee who indicates simpler proofs for Theorem 3.1 and Theorem 5.1. This research is partly supported by Grant-in-Aid for Scientific Research on Priority Areas "Discovery Science" from the Ministry of Education, Science and Culture, Japan.

References

[1] T. Asano, D. Chen, N. Katoh, and T. Tokuyama. Polynomial-time solutions to image segmentations. In *Proc. 7th ACM-SIAM Symposium on Discrete Algorithms*, pages 104–113, 1996.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[3] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Constructing efficient decision trees by using optimized association rules. In *Proceedings of the 22nd VLDB Conference*, pages 146–155, Sept. 1996.

[4] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Data mining using two-dimensional optimized association rules: Scheme, algorithms, and visualization. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, pages 13–23, June 1996.

[5] M. R. Garey and D. S. Johnson. *Computer and Intractability. A Guide to NP-Completeness*. W. H. Freeman, 1979.

[6] N. Katoh. Private communication, Jan. 1997.

[7] C. Lund and M. Yannakakis. On the hardness of approximating minimization problems. *J.ACM*, 41(5):960–981, 1994.

[8] Y. Morimoto, H. Ishii, and S. Morishita. Efficient construction of regression trees with range and region splitting. In *Proceedings of the 23rd VLDB Conference*, pages 166–175, Aug. 1997.

[9] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[10] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[11] J. R. Quinlan and R. L. Rivest. Inferring decision trees using minimum description length principle. *Information and Computation*, 80:227–248, 1989.

[12] K. Yoda, T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Computing optimized rectilinear regions for association rules. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pages 96–103, Aug. 1997.

Appendix

Proof of Theorem 2.1

We first prove the concavity of $Ent(x,y)$. For any $x>y>0$ and any real numbers δ_1 and δ_2, let V denote $\delta_1 x + \delta_2 y$. It suffices to prove $\partial^2 Ent(x,y)/\partial V^2 \leqq 0$. Recall that

$$Ent(x,y) = \frac{x}{n} ent(\frac{y}{x}) + \frac{n-x}{n} ent(\frac{m-y}{n-x}).$$

Define $f(x,y) = \frac{x}{n} ent(\frac{y}{x})$. Then,

$$Ent(x,y) = f(x,y) + f(n-x, m-y) \qquad (1)$$

To prove $\partial^2 Ent(x,y)/\partial V^2 \leqq 0$, it is sufficient to show the following inequalities:

$$\partial^2 f(x,y)/\partial V^2 \leqq 0 \qquad \partial^2 f(n-x, m-y)/\partial V^2 \leqq 0.$$

Here we will prove the former inequality. The latter can be proved in a similar way. When $\delta_1, \delta_2 \neq 0$, we have:

$$f(x,y) = \frac{1}{n}(-y \log \frac{y}{x} - (x-y)\log(1-\frac{y}{x}))$$

$$\begin{aligned}
&\frac{\partial f(x,y)}{\partial V} \\
&= \frac{\partial f(x,y)}{\partial x}\frac{\partial x}{\partial V} + \frac{\partial f(x,y)}{\partial y}\frac{\partial y}{\partial V} \\
&= \frac{1}{n}(-\log(x-y) + \log x)\frac{1}{\delta_1} + \\
&\quad \frac{1}{n}(-\log y + \log(x-y))\frac{1}{\delta_2} \\
&\frac{\partial^2 f(x,y)}{\partial V^2} \\
&= \frac{1}{n}((-\frac{1}{x-y} + \frac{1}{x})\frac{1}{\delta_1} + \frac{1}{x-y}\frac{1}{\delta_2})\frac{1}{\delta_1} + \\
&\quad \frac{1}{n}(\frac{1}{x-y}\frac{1}{\delta_1} + (-\frac{1}{y} - \frac{1}{x-y})\frac{1}{\delta_2})\frac{1}{\delta_2} \\
&= -\frac{1}{n\delta_1^2\delta_2^2 x(x-y)y}(\delta_2 y - \delta_1 x)^2 \\
&\leqq 0 \qquad (x > y > 0)
\end{aligned}$$

When $\delta_1 = 0, \delta_2 \neq 0$, $V = \delta_2 y$, and we have:

$$\begin{aligned}
\frac{\partial f(x,y)}{\partial V} &= \frac{1}{n}(-\log y + \log(x-y))\frac{1}{\delta_2} \\
\frac{\partial^2 f(x,y)}{\partial V^2} &= \frac{1}{n}(-\frac{1}{y} - \frac{1}{x-y})\frac{1}{\delta_2^2} \\
&= \frac{1}{n}\frac{-x}{(x-y)y}\frac{1}{\delta_2^2} \leqq 0 \qquad (x > y > 0)
\end{aligned}$$

The case when $\delta_1 \neq 0$ and $\delta_2 = 0$ can be handled in a similar way.

Next we prove that $Ent(x,y)$ is maximum when $y/x = m/n$. From Equation (1), observe that for any $0 \leqq y \leqq x$, $Ent(x,y) = Ent(n-x, m-y)$. According to the concavity of $Ent(x,y)$, we have

$$\begin{aligned}
Ent(x,y) &= \frac{1}{2}Ent(x,y) + \frac{1}{2}Ent(n-x, m-y) \\
&\leqq Ent(\frac{x+n-x}{2}, \frac{y+m-y}{2}) \\
&= Ent(\frac{n}{2}, \frac{m}{2}),
\end{aligned}$$

which means that $Ent(x,y)$ is maximum when $(x,y) = (\frac{n}{2}, \frac{m}{2})$. Finally we can prove that $Ent(x,y)$ is constant on $y = (m/n)x$, because

$$\begin{aligned}
&Ent(x,y) \\
&= \frac{x}{n}(-\frac{m}{n}\log\frac{m}{n} - (1-\frac{m}{n})\log(1-\frac{m}{n})) + \\
&\quad \frac{n-x}{n}(-\frac{m}{n}\log\frac{m}{n} - (1-\frac{m}{n})\log(1-\frac{m}{n})) \\
&= -\frac{m}{n}\log\frac{m}{n} - (1-\frac{m}{n})\log(1-\frac{m}{n}).
\end{aligned}$$

Since $(\frac{n}{2}, \frac{m}{2})$ is on $y = (m/n)x$, $Ent(x,y)$ is maximum when $y/x = m/n$. ∎

Proof of Theorem 2.2

The interclass variance can be transformed as follows:

$$|R_1|(\mu(R_1) - \mu(R))^2 + |R_2|(\mu(R_2) - \mu(R))^2$$

$$= -|R|\mu(R)^2 + (|R_1|\mu(R_1)^2 + |R_2|\mu(R_2)^2),$$

because $|R| = |R_1| + |R_2|$ and $|R_1|\mu(R_1) + |R_2|\mu(R_2) = |R|\mu(R)$. Since $|R|$ and $\mu(R)$ are constants, the maximization of the interclass variance is equivalent to the maximization of $|R_1|\mu(R_1)^2 + |R_2|\mu(R_2)^2$.

On the other hand, the intraclass variance can be transformed as follows:

$$\frac{\sum_{t\in R_1}(t[A] - \mu(R_1))^2 + \sum_{t\in R_2}(t[A] - \mu(R_2))^2}{|R|}$$
$$= \frac{\sum_{t\in R} t[A]^2 - (|R_1|\mu(R_1)^2 + |R_2|\mu(R_2)^2)}{|R|},$$

because $\sum_{t\in R_1} t[A] = |R_1|\mu(R_1)$ and $\sum_{t\in R_2} t[A] = |R_2|\mu(R_2)$. Since R is fixed, $\sum_{t\in R} t[A]^2$ is a constant. Thus the minimization of the intraclass variance is equivalent to the maximization of $|R_1|\mu(R_1)^2 + |R_2|\mu(R_2)^2$ that is also equivalent to the maximization of the interclass variance. ∎

Proof of Theorem 2.3

The proof is similar to the proof of Theorem 2.1. We first prove that $Var(x, y)$ is a convex function for $0 < x < n$. For any $0 < x < n$ and any δ_1 and δ_2, let V denote $\delta_1 x + \delta_2 y$. Here we will prove the case when $\delta_1, \delta_2 \neq 0$. The other cases can be shown similarly. It is sufficient to prove that $\partial^2 Var(x, y)/\partial V^2 \geqq 0$. Recall that

$$Var(x, y) = x(\frac{y}{x} - \frac{m}{n})^2 + (n - x)(\frac{m - y}{n - x} - \frac{m}{n})^2.$$

Define $g(x, y) = x(\frac{y}{x} - \frac{m}{n})^2$. Then,

$$Var(x, y) = g(x, y) + g(n - x, m - y). \quad (2)$$

We prove $\partial^2 Var(x, y)/\partial V^2 \geqq 0$ by showing the following two inequalities:

$$\partial^2 g(x, y)/\partial V^2 \geqq 0 \qquad \partial^2 g(n - x, m - y)/\partial V^2 \geqq 0.$$

We prove the former case. The latter can be shown in a similar manner.

$$\frac{\partial g(x, y)}{\partial V} = \frac{\partial g(x, y)}{\partial x}\frac{\partial x}{\partial V} + \frac{\partial g(x, y)}{\partial y}\frac{\partial y}{\partial V}$$
$$= \frac{1}{\delta_1}\{(\frac{m}{n})^2 - (\frac{y}{x})^2\} + \frac{2}{\delta_2}(\frac{y}{x} - \frac{m}{n})$$
$$\frac{\partial^2 g(x, y)}{\partial V^2} = \frac{2}{x}(\frac{y}{\delta_1 x} - \frac{1}{\delta_2})^2 \geqq 0$$

Next we prove that $Var(x, y)$ is minimum when $y/x = m/n$. From Equation (2), we have $Var(x, y) = Var(n - x, m - y)$. From the convexity of $Var(x, y)$,

$$Var(x, y) = \frac{1}{2}Var(x, y) + \frac{1}{2}Var(n - x, m - y)$$
$$\geqq Var(\frac{x + n - x}{2}, \frac{y + m - y}{2})$$
$$= Var(\frac{n}{2}, \frac{m}{2}),$$

which implies that $Var(x, y)$ is minimum when $(x, y) = (\frac{n}{2}, \frac{m}{2})$. It is easy to see that $Var(x, y) = 0$ when $y/x = m/n$. Since $(\frac{n}{2}, \frac{m}{2})$ is on $y/x = m/n$, $Var(x, y)$ is minimum when $y/x = m/n$. ∎

Shinichi Morishita Shinichi Morishita is an associate professor of Department of Complexity Science and Engineering, Graduate School of Frontier Sciences, University of Tokyo. His current research interests are database query optimization, web-based database systems, data mining and genome informatics. Email: moris@is.s.u-tokyo.ac.jp

Akihiro Nakaya He is a researcher at Institute of Medical Science, University of Tokyo, and holds BS (1994) and MS (1996) in information science from University of Tokyo.