Discrete Applied Mathematics 3 (1981) 67–72
North-Holland Publishing Company

NOTE

OPTIMAL KEY FOR TAXONS ORDERED IN ACCORDANCE WITH THEIR FREQUENCIES

R.E. KRICHEVSKII, B.Ya. RYABKO and A.Yu. HARITONOV

Institute of Mathematics and Institute of Biology, The Siberian Division of the Academy of Science of the USSR, Novosibirsk, USSR

Received 20 July 1979
Revised 14 April 1980

Assume nothing is known about the probabilities of taxons except that the first taxon is more probable than the second one, the second is more probable than the third, and so on. Then to construct the optimal key, one can regard the ith taxon as having probability proportional to $(i-1)^{i-1}/i^i$.

1. Introduction

Dichotomous keys are widely used in biology, mineralogy, pattern recognition, and other areas. Such a key is labelled complete binary tree. The label of a leaf is the name of a taxon and the label of any other node is an attribute. One goes to the right son of a node if the taxon has the corresponding attribute, otherwise one goes to the left son. Fig. 1 shows a key of dragonflies of the sub-family Libellulinae [1]. The key is based on analysis of the wings of a dragonfly. The first attribute α_1 is equal to 0 if the last antenodal vein of the wing is complete, 1 if it is incomplete. The attribute α_2 is equal to 0 if the sectors of the arculus (an element of the wing) are divided from the very base, and equal to 1 if there is a stem at the top of the sectors, etc.

There are many ways to construct a key, and we want an efficient one. If one wants to reduce the volume of a key to the minimum, one can use the results in [2]. Our goal here is to construct keys with the minimum expected time of identification.

Suppose that we are given n taxons, $n>0$, and let P_i be the probability of occurrence of the ith taxon, such that $P_1+P_2+\cdots+P_n=1$. If the length of a path from the root of a determining tree T up to a leaf corresponding to the ith taxon is $l_i(T)$, then the expected number of attributes needed for identification (expected time) is

$$S(T,P)=\sum_{i=1}^{n} P_i l_i(T).$$

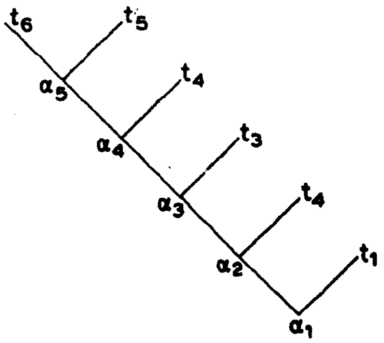

Fig. 1. The key for dragonfly's genera of the subfamily Libellulinae from [1].

In the example of Fig. 1, $l_1(T)=1$, $l_2(T)=2, \dots, l_6(T)=5$. The minimum value of $S(T, P)$ is equal to the entropy

$$H(P) = -\sum_{i=1}^{n} P_i \log_2 P_i.$$

The redundancy of a key T on a vector P is

$$R(T, P) = S(T, P) - H(P).$$

The function $R(T, P)$ is convex. If the probabilities $P_1, \dots, P_n$ are known, then finding the optimal key is equivalent to constructing the optimum decipherable code. The solution is the Huffman code [3], and the Shannon code [4, 5] is close (within 1) of being optimum.

One can scarcely hope to know the exact probabilities. They change from one region to another and from one season to another. Also, it is rather difficult to evaluate the probabilities of a taxon if one has only one or two specimens of it. But it is usually known which taxons occur more frequently than others. We consider the relation $i \leqslant j$ (i occurs less often than j) to be defined for all pairs of taxons and our set of taxons to be linearly ordered with respect to this relation. In the example of Fig. 1, the genus Sympetrum (t_6) is the most numerous (in Siberia), and the genera Leucorrhinia (t_1), Libellula (t_2), Orthetrum (t_4), Pantala (t_3), Neurotemis (t_5) follow in order ($6 \geqslant 1 \geqslant 2 \geqslant 4 \geqslant 3 \geqslant 5$). For a tree T with n leaves we define by $R_n(T)$ the maximum of $R(T, P)$ taken over all the vectors P compatible with a linear order on the set of leaves of T (taxons). We explain in this paper how to make a tree T with n leaves which has the minimum possible value of $R_n(T)$. Our method diminishes $S(T, P)$ by a factor of 1.5 to 2 for most of the keys in [1].

It is possible to generalize thc method to the case of an arbitrary, not necessarily linear, order on the set of taxons. Such a generalization is discussed in [6]. The question is reduced there to finding the capacity of a communication channel.

2. Definitions and auxiliary results

Let a linear order be defined on the set of n taxons, $n>1$: $1 \geqslant 2 \geqslant \cdots \geqslant n$. Let P_n be the set of probability vectors compatible with this order, i.e., P_n is the set of vectors $P=(P_1,\dots,P_n)$ such $P_1 \geqslant P_2 \geqslant \cdots \geqslant P_n \geqslant 0$ and $P_1+P_2+\cdots+P_n=1$. We denote by q^i the n-dimensional vector, the first i coordinates of which are equal to $1/i$ and the last $n-i$ equal to $0, 1 \leqslant i \leqslant n$. Let $Q_n=\{q^1,\dots,q^n\}$.

Lemma 1. *The set P_n is the convex hull of the set Q_n.*

Proof. Let $P=(P_1,\dots,P_n) \in P_n$, $\varphi_i = i\ (P_i-P_{i+1})$, $i=1,\dots,n$. Let $P_{n+1}=0$. Obviously, $\varphi_i \geqslant 0$, $i=1,\dots,n$, $\sum_{i=1}^{n} \varphi_i q^i = P$ and $\sum_{i=1}^{n} \varphi = 1$. □

Lemma 2. *Let* $\lambda_i>0$, $\mu_i>0$, $i=1,\dots,n$; $\lambda_i \geqslant \lambda_{i+1}$, $i=1,\dots,n-1$;

$$\sum_{i=1}^{m} \log \mu_i > \sum_{i=1}^{m} \log \lambda_i, \quad m=1,\dots,n. \tag{1}$$

Then

$$\sum_{i=1}^{n} \mu_i > \sum_{i=1}^{n} \lambda_i. \tag{2}$$

Proof. Let $\varepsilon_i = \mu_i/\lambda_i - 1$. Obviously $\varepsilon_i > -1$. From (1) and the inequality $\ln(1+\varepsilon) \leqslant \varepsilon$, $(\varepsilon > -1)$ we have

$$\sum_{i=1}^{s} \varepsilon_i > 0, \quad s=1,\dots,n. \tag{3}$$

We rewrite (2) in the form

$$\sum_{i=1}^{n} \lambda_i \varepsilon_i > 0. \tag{4}$$

Multiplying the ith inequality (3) by the positive number $\lambda_s - \lambda_{s+1}$, $s=1,\dots,n$, $\lambda_{n+1}=0$, and summing all of them up, we obtain (4). □

3. The method

Suppose that we are given a linearly ordered set of n taxons. To construct an optimum tree, we first of all supply the ith taxon with the probability $\lambda_i = (i-1)^{i-1}/i^i \gamma_n$, where $\gamma_n = \sum_{i=1}^{n} (i-1)^{i-1}/i^i$, $i=1,\dots,n$. It is clear that $\lambda_i > \lambda_{i+1}$, $\sum_{i=1}^{n} \lambda_i = 1$ and

$$\left(\sum_{i=1}^{s} -\log \lambda_i\right) \Big/ s - \log s = \log \gamma_n, \quad s=1,\dots,n \tag{5}$$

Now construct the Shannon code for the probability vector $(\lambda_1, \ldots, \lambda_n)$. The code for the ith taxon is the first $\lceil \log 1/\lambda_i \rceil$ binary digits of the number $\sum_{j=1}^{i} \lambda_j$ $(i = 1, \ldots, n)$. These codes define full binary tree T_M (the nodes with only one son are omitted). It is known that

$$l_i(T_M) \leqslant \log 1/\lambda_i + 1, \tag{6}$$

where $l_i(T_M)$ is the number of attributes required to identify the i-th taxon in the tree T_M.

Theorem. *The maximum redundancy $R_n(T)$ of any tree with n leaves on the linear ordered set of n taxons is at least* $\log_2 \gamma_n$. *On the other hand, for the tree constructed above,*

$$R_n(T_M) \leqslant \log_2 \gamma_n + 1. \tag{7}$$

In other words, the tree T_M is optimum to within the additive constant 1.

Proof. Let T be a tree with n leaves, and let $\mu_i = 2^{-l_i(T)}$, $i = 1, \ldots, n$. The Kraft inequality is valid:

$$\sum_{i=1}^{n} \mu_i = 1. \tag{8}$$

If $R_n(T) < \log_2 \gamma_n$, then $R(T, q^s) < \log_2 \gamma_n$, $s = 1, \ldots, n$. The last inequality is equivalent to

$$\sum_{i=1}^{n} q_i^s(l_i(T) + \log_2 q_i^s) < \log_2 \gamma_n.$$

Hence we have

$$\left(-\sum_{i=1}^{s} \log_2 \mu_i\right) \Big/ s - \log_2 s < \log_2 \gamma_n, \quad s = 1, \ldots, n. \tag{9}$$

Lemma 2, (5), (8) and (9) yield

$$1 = \sum_{i=1}^{n} \mu_i > \sum_{i=1}^{n} \lambda_i = 1,$$

a contradiction. Hence $R_n(T) \geqslant \log_2 \gamma_n$. The first statement of the theorem is now proved.

The convex function $R(T, P)$ attains its minimum on a vector of Q_n (Lemma 1). From (6) and (9), we have $R(T_M, q^s) \leqslant \log_2 \gamma_n + 1$, $s = 1, \ldots, n$ and (7) is proved. □

It is possible to use the Fano or Huffman codes instead of the Shannon code. Fig. 2 presents the optimum trees for the ordered sets with not more than 9 taxons.

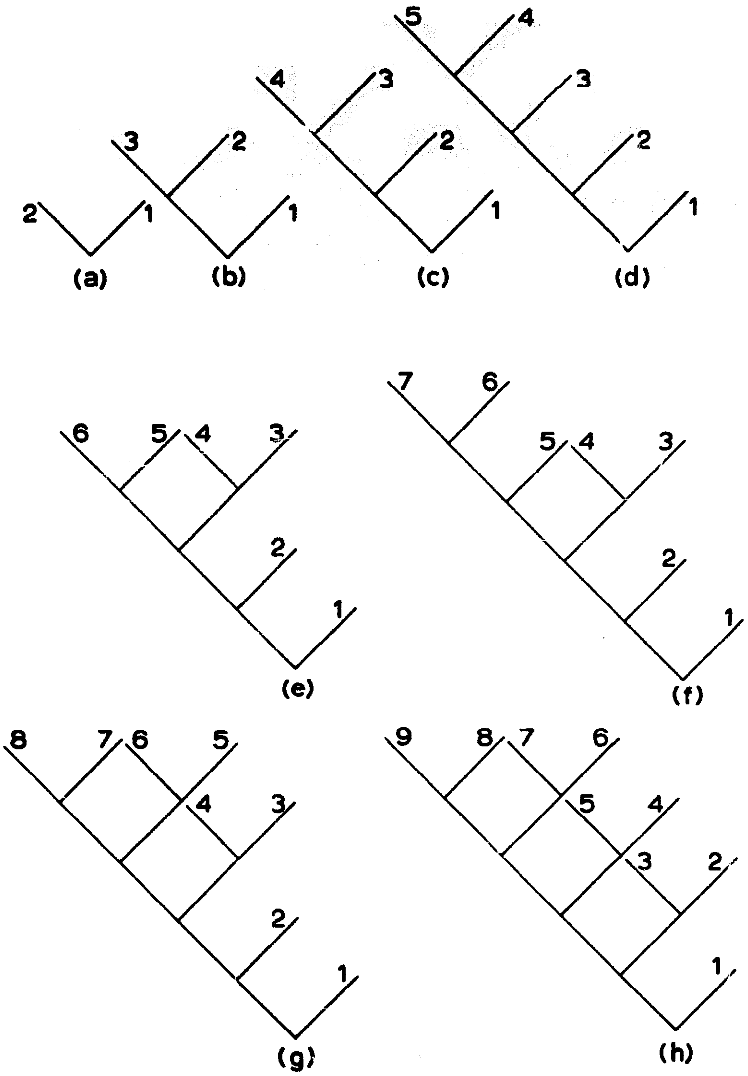

Fig. 2. The optimal keys.

4. An application

We took several sets of taxons from [1] and constructed keys for them by our method. For example, we could find attributes for the family Libellulinae in order to give it the key of Fig. 2(e). To compare the new and the former determining trees, one has to know the probabilities, which we did not know. We decided to count the number of attributes necessary to determine taxons in real samples. For example, B.F. Belyshev, making an inventory of Far East dragonflies, possessed a collection of 226 specimens of Libellulinae [7]. This collection included all the specimens gathered by many researchers for a long time. There are reasons to consider this collection to be a representative sample. 146 of these dragonflies are classified as Sympetrum, 69 as Leucorrhinia, 5 as Libellula, 6 as Pantala. To determine Sympetrum with the determiny tree in 1, Fig. 1, one has to use 5 attributes, Leucorrhinia 2, Libellula 3, Pantala 4. The total number of attributes needed to determine all the dragonflies of this collection is

$$146\times 5+69\times 2+5\times 3+6\times 4=907.$$

This amount is reduced to

$$146\times 1+69\times 2+5\times 4+6\times 4=328$$

attributes with the help of our determination tree 2e. We also made calculations of this kind for the data from [7, 8, 9]. The results were good: the use of our method often diminished the expected number of attributes by a factor of between 1.5 and 2.

Acknowledgement

It is a great pleasure to express our gratitude to Mr. P.L. Hammer for his kind help in preparing this paper.

References

[1] B.F. Belyshev, Key of dragonflies of Siberia by imaginal and larval stages, M–L 1963 113 pp. (in Russian).

[2] R.E. Krichevskii, Digital enumeration of binary dictionaries, Soviet Math. Dokl. 19(2) (1978) 469–473.

[3] R.M. Fano, Transmission of Information: A Statistical Theory of Communications (M.I.T. Press and J. Wiley, New York and London, 1961).

[4] L.E. Morse, Computer programs for specimen identification, key construction and discription using taxonomic data matrices, Publ. Mus. Michigan State Univ. (Biol. Ser.) 5 (1) (1974).

[5] R.J. Pankhurst, Automated identification in systematics, Taxon 23 (1) (1974) 45–51.

[6] B.Ya. Ryabko, Encoding of a source with unknown but ordered probabilities, in: Problems of Transmission of Information, No. 2 (1979) 71–77 (in Russian).

[7] B.F. Belyshev, Materials for cognition of dalnevostochnoi fauna Odonata, Trudi Dal. Vost. filial AN SSSR 3 (6) (Wladivostok, 1956) 181–199 (in Russian).

[8] B.F. Belyshev and T.N. Gagina, On the fauna of Odonata of the Baikal Region, Fragm. faun. 7 (9) (Warszawa, 1959) 159–178 (in Russian).

[9] B.F. Belyshev, H. Remm and A.G. Pankratyev, On the odonatological fauna of Ussuri Territory, Zhiv. Prir. Dal. Vost. (Tallin, 1971) 162–170 (in Russian).