Numerical Approximations to Expectations of Functions of Binary Sequences Subject to Error

by

D M Jackson

and

L J White*

70-82

Technical Report

NO. 70-82

November, 1970

Department of Computer Science
Cornell University
Ithaca, New York 14850

*Department of Computer and
Information Science
Ohio State University
Columbus, Ohio 43210

This work was supported in part by National Science Foundation Grants GN-27224 and GN-534.1.

Numerical Approximations to Expectations of Functions of Binary Sequences Subject to Error

There is growing interest in devising non-statistical classification algorithms for multivariate populations. Statistical algorithms are avoided either because they are too costly, or because an adequate statistical model for the population does not exist (e.g., use of trainable linear machines in pattern recognition). Such algorithms may be sensitive (unstable) to errors in their data. The particular case of populations of objects characterised by binary attributes susceptible to independent and equiprobable errors is examined. The determination of stability requires the prior computation of the expectation of a statistical function of the object-pair similarities. The order and convergence of a numerical approximation for determining these expectations with prescribed accuracy is examined in the sub-asymptotic case in which normality does not occur. A number of results are given.

1. Introduction

In recent years, growing attention has been given to providing techniques for classifying populations whose underlying statistical structure is unknown. In a typical situation, a multivariate heterogeneous population of objects is given together with a characterization of each object in terms of observable attributes. The objective of the analysis is to determine a decomposition of the population which will reveal the population's structure and which will suggest further models (1), for a more sophisticated treatment of the material.

Briefly, non-statistical classification involves the use of similarity functions and classificatory criterion functions. A similarity function specifies the rule for computing resemblence between arbitrary pairs of objects (2). A criterion function determines whether a particular collocation of objects is admissible as a class of the classification (3,4,5, inter al.). We shall be concerned here with the stability requirement, namely that a classification must not change appreciabley with respect to small changes in the material to be classified. Instability indicates that the objects have been inadequately characterized. To facilitate the investigation of stability, subsequent analysis provides expressions for the expectation $\emptyset^*$ of statistical functions $\emptyset$ (mean, variance, moments, factorial moments) of the similarity f between arbitrary pairs of objects when the states of the N attributes are susceptible to error independently and equiprobably with probability r. A low order technique for computing the error matrix, $E^{(N)}$, formed by the expectations for each pair of objects will be discussed. The use of the error matrix and the quantification of small changes in a classification have been discussed elsewhere (6,7) and will not be given here.

2. Algebraic specification of the error matrix

It may be shown using the central limit theorem that for large N and n_1, n_2, n_3, comparable in magnitude, then $\emptyset^*$ is asymptotically approximately normal. However, we are concerned here with determining $\emptyset^*$ with prescribed accuracy for intermediate values of N for which normality does not occur.

Suppose that a multivariate heterogeneous population consists of M objects $\{t_1\}$ and that the characterizations of these objects are provided by composition from a population of N binary attributes. Denote by a the (unordered) pair of attribute states (0,0), where for some attributes a_i, t_j has state 0 and t_k has state 0. Similarly, denote by x and d the paired combinations of states (1,0)V(0,1) and (1,1) respectively. For the pair of objects (t_j, t_k), let n_1, n_2, n_3 be numbers of attributes having the combinations of states a,x,d respectively. For brevity, denote (n_1, n_2, n_3) by $\underline{n}$. The characterizations are scrambled by errors into different ones. Let m_1, m_2, m_3 and $\underline{m}$ be the analogous quantities developed from these. P(u,v) is the probability of the transition v→u, where u and v are both one of a, x, d. The following two forms for $\emptyset^*(\underline{n})$ will be used extensively (15):

1)
$$\emptyset^*(\underline{n}) = S^{2N} \sum_{\underline{m}} h(\underline{m},\underline{n}).\emptyset(\underline{m},\underline{n})$$

where i) $h(m,n) = H_{\underline{m}}(\{t^2,2t,1\}^{n_1}.\{t,1+t^2,t\}^{n_2}.\{1,2t,t^2\}^{n_3})$

ii) $\{x_1,x_2,x_3\} = (x_1,x_2,x_3).(q_1,q_2,q_3)$

iii) $H_{\underline{m}}(X) =$ coeff. of $\prod_{i=1}^{3} q_i^{m_i}$ in X where X is a polynomial in the q_i's.

$$2)\qquad \phi^*(\underline{n}) = S^{2N} \sum_{i=0}^{2(n_1+n_3)+n_2} t^i \sum_{\substack{I\in I^*(\underline{n}) \\ k(I)=i}} (1+t^2)^{I_{22}} \cdot K_I \cdot \phi(\underline{m},\underline{n})$$

where i) $I^*(\underline{n})$ is the set of all tables such that $\sum_{i=1}^{3} I_{ij}=n_j$

ii) $k(I) = N - \underline{\text{trace}}(I)+I_{13}+I_{31}$ (table constant)

iii) $m_i = \sum_{j=1}^{3} I_{ij}$ and $\sum_{i=1}^{3} m_i=N$ where $N = \sum_{j=1}^{3} n_j$

iv) $K_I=2^{m_2-I_{22}} \prod_{i=1}^{3} M(n_i:I_{i1},I_{i2})$ using the usual multinomial

for $$\frac{n_i!}{I_{i1}!\ I_{i2}!\ (n_i-I_{i1}-I_{i2})!}$$

In both (1) and (2), $t = r/s$ and $s = 1-r$.

A final form might also be mentioned(16):

$$3)\qquad \phi^*(\underline{n}) = \sum_{\underline{i}} a_{\underline{i}} . B_{\underline{i}}$$

where i) $\phi(\underline{m},\underline{n}) = a_{\underline{i}}(\underline{n}) . \underline{m}^{\underline{i}}$

ii) $D_{ij} = P(i,j) . \frac{d}{dP(i,j)}$ is a formal differential operator.

iii) $B_{\underline{i}} = (\prod_{j}(\sum_{k} D_{jk})^{i_j}) . (\prod_{k}(\sum_{j} P(j,k))^{n_k})$

It is anticipated that this expression may be expanded quite readily by a suitable algebra system (8,9) to give an analytic solution for each element of the error matrix.

In a more general context, the above formulation applies to problems encountered in the comparison of pairs of fixed length binary sequences, and is suggestive of the classical Matching

Problem (10,11). $\emptyset^*$ may be regarded as the expectation of a generalized 'match' between two arbitrary compositions of objects of two kinds (labelled 0, and 1) when there is a probability of misrecognizing the kind of an object. The third form of $\emptyset^*$ results from a method related to the symbolic method (12), and its extensions (13).

Chief attention for computational purposes is confined to $\emptyset = |f(\underline{n})-f(\underline{m})|$ where $f(\underline{n})$ is the Jaccard similarity function $n_3/(n_2+n_3)$. The relevance of the former to the stability of classifications has been discussed elsewhere (14).

3. Number of Tables

Evaluation of $\emptyset^*(\underline{n})$ involves summation over all feasible tables, I, such that the row sums are given by $\underline{n}$. Bell (15) has computed the number of matrices for the special case of generalized stochastic matrices. The number of possible vectors, $\underline{n}$, is M(N+2:2), and the number of feasible tables, T(N), for given N is M(N+8:8). If $T(\underline{n},k)$ is the number of tables with specified table constant k, and $T^*(\underline{n},k)$ is the value of $T(\underline{n},k)$ in the case $k \leq \min(n_1,n_2,n_3)$ then:

$$T^*(\underline{n},k) = (k+2)(k+4)(k+6)(k^2+8k+10)/2.4.6.10 \quad \text{if k even}$$

$$= (k+1)(k+3)(k+4)(k+5)(k+7)/2.4.6.10 \quad \text{if k odd}$$

If $R(N,\bar{k})$ gives the proportion of valid tables with $k \leq \bar{k}$ to tables with k unrestricted then:

$R^*(N,\bar{k}) = M(N+2:2)(M(N+8:8))^{-1}.(T^*(\underline{n},1)+T^*(\underline{n},2)+\ldots+T^*(\underline{n},\bar{k}))$ is an upper bound for $R(N,\bar{k})$. We can show that $R^*(N,\bar{k})$ is asymptotic in N to $R(N,\bar{k})$ for $\bar{k} << N$, and large N. Necessary and sufficient conditions for the existence for a table are given by Jackson and White (15).

4. Experimental Results

The computation time required to calculate $\emptyset^*(\underline{n})$ depends on the number of tables to be considered. The latter increases on the

order of N^8. Much of the work in determining an efficient method centres around reducing the number of tables utilized while yielding $\emptyset^*$ to a prescribed accuracy. All computations were carried out on an IBM 360/75.

Table I shows several cells of the matrix $E^{(10)}$ for successive values of $\bar{k}$ and r=0.05. It is seen that sufficient accuracy is given by $\bar{k}$=5. A proof of uniform convergence for all $\underline{n}$ for the same $\bar{k}$ is given in Section 6, and it is clear that $\emptyset^*$ may be approximated by limiting the value of $k<<\bar{k}$. The behavior of $\bar{k}$ with respect to N and r is given in Table II.

Table III indicates the reduction in computation time obtained by restricting k for several values of $\bar{k}$ with r=0.05. Only tables with $k\leq\bar{k}$ are constructed. Column four gives the ratio of the computation times with $k\leq\bar{k}$ and k unrestricted. The ratio of valid tables to the total number of tables provides a good estimate of the time saving ratio, and is proposed for this use. Column six gives an indication of the accuracy of the truncated calculation by specifying the error in the third significant figure.

For N=100 the matrix element corresponding to $\underline{n}$=(99,0,1) for $\bar{k}$=21 and 22, was estimated as 0.883 and 0.884 respectively. $\bar{k}$=21 gave sufficient accuracy and took approximately 0.15 seconds. The complete matrix took approximately 15 minutes. Without table truncation, the time is estimated to be approximately three years.

5. Monotonicity

The observed behavior for $E^{(10)}$, given in Table IV, has been noted for all other values of N and r, and is accounted for by the following theorem:

Theorem 5.1 For $t\varepsilon(0,t_0)$ for some t_0 then:

i) $\phi^*(\underline{n})$ is strictly monotonically increasing wrt n_1 and $n_3 (n_3 \geq 1)$, with n_1 and n_3 as independent variables.

ii) $\phi^*(n_1,0,n_3)$ is strictly monotonically increasing wrt $n_1 (n_3 \geq 1)$

iii) $\phi^*(N-1,0,1) > \phi^*(N,0,0)$

Proof:

i) $S^{-2N}\phi^*(\underline{n})$ is a polynomial of degree 2N in t, so $S^{-2N}\phi^*(\underline{n})$ is dominated by its constant and linear term in t, for t sufficiently small, i.e. $t \varepsilon (0,t_0)$. Let $X=n_2+n_3>0$. The case $n_2=n_3=0$ is excluded for special treatment since $f(\underline{n})$ is undefined for (N,0,0). For $n_2 \geq 1$, then

$$S^{-2N}(\phi^*(n_1,n_2-1,n_3+1)-\phi^*(n_1,n_2,n_3)) = X^{-1}(2n_1(X+1)^{-1}+2(n_2+1)(X-1)^{-1}+1)t$$

so $\phi^*(\underline{n})$ is strictly monotonically increasing wrt n_3.

Similarly:

$$S^{-2N}(\phi^*(n_1,n_2,n_3)-\phi^*(n_1-1,n_2+1,n_3) = 2n_3((X-1)(X)(X+1)(X+2))^{-1}.$$

$$(2n_1(X-1)+n_2(X+2)+X(X-1))t+0(t^2)$$

so $\phi^*(\underline{n})$ is strictly monotonically increasing wrt n_1 for $n_1 \geq 1$ and $n_3 \geq 1$

ii) If $n_3 \geq 1$ then $\phi^*(n_1,0,n_3) = 2(1+n_1(n_3+1)^{-1})t+0(t^2)$ and is strictly monotonically increasing wrt n_1.

iii) $\phi^*(N-1,0,1)=(N+1)t+0(t^2)$ and $\phi^*(N,0,0)=0(t^2)$ so

$$\phi^*(N-1,0,1)>\phi^*(N,0,0)$$

For particular elements, the value of t_0 may be estimated by the following theorem:

Theorem 5.2 If $t\in(0,N^{-1/2})$ and $\phi(\underline{m},\underline{n})\geq 0$ for all $\underline{m}$, $\underline{n}$ then

$\phi^*(n_1,0,n_3)$ is a monotonically increasing function of n_1.

Proof:

From 2.1 we know that $\phi^*(n_1,0,n_3)-\phi^*(n_1-1,0,n_3+1)=S^{2N}(1-t^2)$.

$$\sum_{\underline{m}} \phi(\underline{m},\underline{n}).H_{\underline{m}}(\{-1,0,+1\}.\{1,2t,t^2\}^{n_3}.\{t^2,2t,1\}^{n_1^{-1}})$$

$$=S^{2(N-1)}.(1-t^2).\sum_{\underline{m}} \phi^*(\underline{m},\underline{n}).\ t^{2\,m_1+m_2}.\sum_{\underline{j}} M(n_1-1:\underline{j}).t^{2(j_3-j_1)}.$$

$$\left(-(m_1-j_1)+t^{-2}(m_3-j_3)\right)/(n_3+1)\Big).M(n_3+1:\underline{m}-\underline{j})$$

≥ 0 if $t^2\leq \min_{\underline{m}}((m_1-j_1)/(m_3-j_3))$ where $\underline{j}<\underline{m}$ (component by component)

$= 1/N$ and theorem follows.

We have assumed that $m_1, m_3, m_1-j_1, m_3-j_3$ are non-zero in certain cases. We may show that the terms for which some of these are zero do not affect the conclusion. This is omitted for brevity.

6. Convergence

If f(N,0,0) is defined to be unity, then this is the largest element in the matrix. The tables of results, however, have been computed with f(N,0,0)=0. (N,0,0) occurs only when the two binary sequences consist of zeroes alone. For a proper characterization of the objects, this case will not occur.

The following theorem confirms convergence of the algorithm for the special case $\emptyset^*(N,0,0)$. This is the simplest case and allows extension in Theorems 6.2 and 6.3 to arbitrary n. Theorem 6.1 specifies for the numerical algorithm the summands which may be neglected at a cost of at most ε in the truncation error. We shall first need the following lemma.

Lemma 6.1 If $P_k = \sum_{I\varepsilon I^*(N,0,0),\ k(I)=k} K_I$ and $R_{k+1} = P_{k+1}/P_k$ then:

i) R_k is monotonic decreasing in k and P_k is concave unimodal in k.

ii) P_k has a maximum at k=N and is symmetrical about k=N.

Proof

If $k \leq N$ then $P_k = \sum_{a=0}^{\lfloor \frac{1}{2}k \rfloor} T(a,k)$ and if $k>N$ then $P_k = \sum_{a=k-N}^{\lfloor \frac{1}{2}k \rfloor}$, where $\lfloor X \rfloor$ denotes the integer part of X, and $T(a,k)=2^{k-2a}.M(N:k-2a,a)$.

Suppose that k is even. Then $P_k = \sum_{a=0}^{\frac{1}{2}k} T(a,k)$ and $P_{k+1} = \sum_{a=0}^{\frac{1}{2}k} b(N,k,a).T(a,k)$ where $b(N,k,a) = 2(N-k+a)/(k+1-a)$. Now b(N,k,a) is strictly monotonically decreasing in k, so R_k is also. It follows that P_k is concave unimodal with k. An analogous argument holds for k odd. (ii) follows directly from the expressions for

P_k given above, by change of variable.

Theorem 6.1 Suppose that ϕ^* is non-negative and bounded (without loss of generality assume $0 \leq \phi \leq 1$). Let ϕ^*_i be the partial sum of $\phi^*(N,0,0)$ for $k(I) \leq i$.

Then $\phi^*_{\bar{k}} - \phi^*_{\bar{k}-1} < \delta$ and $t^{\bar{k}}.P_{\bar{k}+1} < t^{\bar{k}}.P_{\bar{k}} \Rightarrow \phi^*_{2N} - \phi^*_{\bar{k}} < \varepsilon$

where $\varepsilon = \begin{cases} \delta(N-\bar{k}+(1-t)^{-1}) & \bar{k}<N \\ \delta t/(1-t) & \bar{k} \geq N \end{cases}$

Proof Let $P'_k = \sum_{\substack{I \in I^*(N,0,0) \\ k(I)=k}} K_I.\phi_I$. Then from the form of ϕ^* given in 2.2 we have $K_I = 2^{I_{12}}.M(N:I_{11},I_{12})$ and $k(I)=I_{12}+2I_{13}$. Now $P'_k \leq P_k$.

as $0 \leq \phi \leq 1$.

Since P_k is concave unimodal in k (from Lemma 6.1) then $t^k.P_k$ is concave unimodal for $t<1$.

Let k_0 be the smallest k such that $t^{k+1}.P_{k+1} < t^k.P_k$. k_0 exists since P_k is unimodal.

k_0 must be the smallest k such that $t.R_k < 1$. Then for any $k > k_0$

$$t^{k+1}.P_{k+1} < t^k.P_k.$$

For $\bar{k}<N$:
$$\phi^*_{2N} - \phi^*_{\bar{k}} = s^{2N} . \sum_{k=\bar{k}+1}^{2N} t^k.P'_k$$
$$\leq s^{2N} . \sum_{k=\bar{k}+1}^{N} t^k.P_k + s^{2N} . \sum_{k=1}^{N} t^k.P_{N+k}$$

(by symmetry of P_k (from Lemma 6.1, and $P'_k \leq P_k$)

$< \delta(N-\bar{k}+(1-t)^{-1})$

For $\bar{k} \geq N$: $\phi^*_{2N} - \phi^*_{\bar{k}} = s^{2N} \sum_{k=\bar{k}+1}^{2N} t^k . P'_k < s^{2N} . \sum_{k=\bar{k}+1}^{2N} t^k . P_{\bar{k}}$

(since $\bar{k} \geq K_0$, $t<1$, P_k monotonic decreasing for $k \geq N$)

$< \delta\, t/(1-t)$

This completes the proof.

Convergence is now shown for any $\phi^*(\underline{n})$ in the matrix $E^{(N)}$, for the values of $\bar{k}$ and δ specified in Theorem 6.1, with truncation error less than the ε of Theorem 6.1.

Theorem 6.2 If $0 \leq \phi \leq 1$ and ϕ^*_i is the partial sum of $\phi^*(n_1,0,n_3)$ for $k(I) \leq i$, then for $\bar{k}$ and δ defined in Theorem 6.1:

$\phi^*_{2N} - \phi^*_{\bar{k}} < \varepsilon$ where ε is defined in Theorem 6.1.

Proof From the form of ϕ^* given in 2.2:

$\sum_{i=0}^{2N} t^i . P_i(n_1,0,n_3) = (1+t)^{2N}$ (weights of the probability distribution)

where $P_i(\underline{n}) = \sum_{\substack{I \varepsilon I^*(\underline{n}) \\ k(I)=i}} K_I$. Since this is an identity for all t, and K_I is independent of t, then:

$P_i(n_1,0,n_3) = M(2N:i)$ which is independent of n_1 and n_3.

Then, using δ and $\bar{k}$ defined in Theorem 6.1, we have:

$$\phi^*_{2N}(n_1,0,n_3) - \phi^*_{\bar{k}}(n_1,0,n_3) = s^{2N} \cdot \sum_{i=k+1}^{2N} t^i . P_i(n_1,0,n_3)$$

$$= s^{2N} \cdot \sum_{i=\bar{k}+1}^{2N} t^i . P_i(N,0,0) < \varepsilon \text{ from Theorem 6.1}$$

This completes the proof.

The following theorem demonstrates convergence for $\phi^*(\underline{n})$. However, for the $\bar{k}$ and δ of Theorem 6.1, the truncation error is larger than the ε given in Theorem 6.1.

Theorem 6.3 If $0 \leq \phi \leq 1$, and ϕ^*_i is the partial sum of $\phi^*(n_1,n_2,n_3)$ for $k(I) \leq i$, then for $\bar{k}$ and δ defined in Theorem 6.1:

$$\phi^*_{2N} - \phi^*_{\bar{k}} < \varepsilon' \quad \text{where } \varepsilon' = (1+t^2)^{n_2} \varepsilon \text{ , } \varepsilon \text{ being defined in Theorem 6.1}$$

Proof Analogous to proof of Theorem 6.2. This inequality $P_i(n_1,n_2,n_3) < P_i(N,0,0)$ is used. This may be shown by a judicious choice of parametrizations for the sets of tables $I^*(n_1,n_2,n_3)$ and $I^*(N,0,0)$.

7. The diagonal elements of an error matrix

A number of interesting results may be proved for the diagonal elements of an error matrix. The diagonal elements are given by $\underline{n}=(n_1,0,n_3)$. We know from Theorem 5.1 that diagonal elements dominate elements in the same row or column and that they are monotonic in n_1 and n_3.

Theorem 7.1 If $\phi(\underline{m},\underline{n})$ is independent of $\underline{n}$ for $\underline{n}=(n_1 0,n_3)$, then

$$\phi^*(n_1,0,n_3:t) = \phi^*(n_3,0,n_1;\tfrac{1}{t})$$

Proof From the form of ϕ^* given in 2.1:

$$\phi^*(n_1,0,n_3)=s^{2N}.\ \sum_{\underline{m}}\ \phi(\underline{m},\underline{n}).H_{\underline{m}}(\{t^2,2t,1\}^{n_1}.\{1,2t,t^2\}^{n_3})$$

$$=(t/(1+t)).\ \sum_{\underline{m}}\ \phi^*(\underline{m},\underline{n}).H_{\underline{m}}(\{t^{-2}.2t^{-1},1\}^{n_1}.\{1,2t^{-1},t^{-2}\}^{n_3})$$

$=\phi(n_3,0,n_1;t^{-1})$ (since $\phi(\underline{m},\underline{n})$ is independent of $\underline{n}$) This completes the proof.

A stronger form of this may be stated for $h(\underline{m},\underline{n})$ and any (n_1,n_2,n_3), namely $h(\underline{m}(n_1,n_2,n_3);t) = h(\underline{m}(n_3,n_2,n_1);t^{-1})$. The proof of this is analogous to the above.

The following identities hold:

Theorem 7.2 i) If f is the Jaccard similarity function then:

$$\sum_{\underline{n}}\ M(N:n_1).\phi^*(n_1,0,n_3)=(4t/(1+4t+t^2)).2N$$

ii) If f is the Jaccard similarity function then:

$$\sum_{\underline{n}}\ M(N:n_1).(-1)^{n_1}.\phi(n_1,0,n_3)=0$$

Proof i) follows straightforwardly from the form of ϕ^* given in 2.1.

ii) follows since the left hand side of the identity is equal to

$$\sum_{\underline{m}}\ (m_2/(m_2+m_3)).H_{\underline{m}}(\{1,0,-1\}^N).(1-t^2)^N.s^{2N}=0,$$

since, if $m_2\neq 0$ then $H_{\underline{m}}(\{1,0,-1\}^N)=0$

Analogous results may be obtained by the same methods for other similarity functions. Particularly simple expressions may be found for a number of the error matrix elements. These are presented in the following theorem.

Theorem 7.3 If f is the Jaccard similarity function then:

i) a) $\phi^*(N,0,0) = (t/(2+t)).(1-s^{2N})$

b) $\phi^*(N,0,0) \to t/(2+t)$ as $N \to \infty$

ii) a) $\phi^*(N-1,0,1) = (2/(2+t)).(1-t^2.s^{2N}-((1-t^2)/(Nt(2+t))).(1-s^{2N}))$

b) $\phi^*(N-1,o,1) \to 2/(2+t)$ as $N \to \infty$

Proof Straightforwardly, from the form for ϕ^* given in 2.1.

Theorem 7.1 may be applied to Theorem 7.3 (ii)(a) to give a closed form expression for $\phi^*(1,0,N-1)$. Summation techniques used in Theorem 7.3 may be used to show that diagonal elements $\phi^*(n_1,0,n_3)$ may be computed by an algorithm of order at most N^2.

Theorem 7.4 If t=1 then:

$$\phi^*(n_1,0,n_3) = (2/3).(1-(\tfrac{1}{3})^{2N})$$

Proof Follows directly from the form for ϕ^* given in 2.1. This has been used for validating the program for computing $\phi^*(\underline{n})$. Table V shows $\phi^*(n_1,0,n_3)$ computed for N=10, t=1. With $\delta=10^{-3}$, $\bar{k}$ had the value 16.

8. Improved estimate for the truncation error.

From Section 6 we know that $T(a+1,k)/T(a,k)$ is monotonically decreasing so $T(a,k)$ is unimodal in a. The dominant term of P_k is given by the smallest a such that $T(a,k) \geq T(a+1,k)$. i.e. $4(a+1)N+6a+4 \geq k(k+3)$. Suppose that a' is such a value and that $\bar{a}=a'+1$. Corresponding terms of P_k and P_{k+1} are related by the factor $b(N,k,a)$, so R_k may be approximated by $\bar{R}_k = b(N,k,\bar{a}) = (4N(N-k)+6N+(k-1)(k-2))/(k+1)(2N-k+1)$. Table VI gives several values of $\bar{R}_k$ for N=10 and $k \leq 10$, and shows that R_k is monotonic decreasing in k, as proved in Lemma 6.1.

$\bar{R}_k$ can be used to obtain a larger value of $\delta(\varepsilon)$ than provided by Theorem 6.1. Clearly, δ is related inversely to computation time. We show that ε and δ are of the same order of magnitude.

Now $$\sum_{k=\bar{k}+1}^{N} t^k . P_k < \delta t.\bar{R}_{\bar{k}}/(1-t.\bar{R}_{\bar{k}}) \quad \text{(by monotonicity of } R_k\text{; Lemma 6.1)}$$

Also $$t^N . \sum_{k=1}^{N} t^k . P_{N-k} < \delta\, t^2/(1-t^2)+\delta t^{2(N-\bar{k})}/(1-t)$$

Thus $$\varepsilon/\delta = t.\bar{R}_{\bar{k}}/(1-\bar{R}_{\bar{k}}.t)+t^2/(1-t^2)+t^{2(N-\bar{k})}/(1-t) \quad \text{if } \bar{k}<N$$

$$= t/(1-t) \quad \text{if } \bar{k}\geq N$$

Table VII gives an improved truncation error estimate over the one given by Theorem 6.1. ε and δ are of the same order of magnitude. The truncation error is therefore estimated by the last summand of ϕ^* which has been included.

9. Conclusions

The algorithm which has been examined has been verified for the Jaccard similarity function. Uniform convergence of the method, however, holds for any statistical function ϕ and similarity function f such that $\phi(\underline{m},\underline{n})$ is bounded and non-negative. All elements of an error matrix converge for the same value of $\bar{k}$. Moreover, it is seen from the computational results that $\bar{k}<<N$. Particular kinds of results, for example monotonicity properties and closed forms, obtained for the Jaccard function are derivable for other similarity functions.

A conservative upper bound on the order of the algorithm is $N^2\bar{k}^6$. Experimental evidence shows that the actual order of the algorithm is considerably less than this. A better estimate of order is under investigation. For a given set of data, not all of the $\underline{n}$ need to occur, and additional economy may be achieved by computing

only the elements of the matrix which are required.

The analysis has provided a basis for determining the stability of classifications of objects described by binary attributes which are susceptible to error independently and equiprobably. The analysis is particularly valuable when the number of attributes is such that the asymptotic normality of $\emptyset^*$ does not hold.

Acknowledgement

This work was supported in part by the National Science Foundation Grants GN-27224 and GN-534.1.

References

1. Nagy, G State of the art in pattern recognition. Proceedings of IEEE, 56, 1968, pp 836-862.

2. Sokal, R R and Sneath, P H A Principles of Numerical Taxonomy. Freeman, San Fransisco, 1963.

3. Lerman, I C On two criteria of classification. In: Numerical Taxonomy (Cole, A J ed) Academic Press, 1969, pp 114-127.

4. Friedman, H P and Rubin J On some invariant criteria for grouping data. Journal of the American Statistical Association, 62, 1967, pp 1159-1178.

5. Parker-Rhodes, A F and Needham, R M The Theory of Clumps. Report no. 126. Cambridge Language Research Unit, Cambridge, England, 1960.

6. Jackson, D M Comparison of classifications. In: Numerical Taxonomy, (Cole A J Ed) Academic Press, 1969, pp 91-111.

7. Jackson D M The Stability of classifications of binary attribute data. Classification Society Bulletin,(in press) 1970.

8. Collins, G E PM A system for polynomial manipulation. Comm. A C M, 9, 1966, pp 578-589.

9. Collins, G E The SAC-1 system: An introduction and survey,*

10. Bernouilli, N Montmort's "Essay d'Analyse sur les Jeux des Hazards". 2nd ed, Paris 1714.

11. De Moivre, A Doctrine of Chances 1st ed, London, 1718

12. McMahon, P A Combinatory Analysis, Vol 1, Ch 3, Cambridge University Press, 1915.

13. Kaplansky, I and Riordan, J. Multiple matching and runs by the symbolic method. Annals of Mathematical Statistics, 16, 1945, pp 272-277.

14. Bell, A G Partitioning the integers in n dimensions, Computer Journal, 13, 1970, pp 278-283.

15. Jackson, D M and White, L J, The weakening of taxonomic inferences by homological errors. Mathematical Biosciences, (in press), 1970.

16. Jackson, D M An error analysis for functions of qualitative attributes with application to information retrieval. Proceeding of the symposium on Computer and Information Science, (Tou, J T ed),3,1969, (in press).

*To be presented at the Second Symposium on Symbolic and Algebraic

n_1	n_3	k=1	k=3	k=5	k=6	k=20(exact)
0	10	0.038	0.090	0.095	0.095	0.095
0	6	0.036	0.062	0.064	0.064	0.064
0	0	0.019	0.047	0.050	0.050	0.050
2	8	0.046	0.108	0.114	0.114	0.114
2	4	0.038	0.067	0.069	0.069	0.069
2	0	0.019	0.047	0.049	0.049	0.049
4	6	0.059	0.135	0.143	0.143	0.143
4	0	0.019	0.046	0.048	0.048	0.048
6	4	0.083	0.187	0.190	0.190	0.190
6	0	0.019	0.045	0.047	0.047	0.047
8	2	0.138	0.278	0.289	0.289	0.289
8	0	0.019	0.042	0.044	0.044	0.044
9	1	0.208	0.325	0.397	0.397	0.397(largest elem
10	0	0.000	0.015	0.016	0.016	0.016

TABLE I Values of $\underset{\bar{k}}{\phi^*}(\underline{n})$ for N=10, r=0.05 and several values of $\bar{k}$.

r	N	$\bar{k}$	r	N	$\bar{k}$
0.05	10	5	0.20	10	10
	20	7		20	16
	30	9		30	21
	40	11		40	27
	50	12		50	32
0.10	10	7	0.25	10	11
	20	10		20	18
	30	14		30	25
	40	17		40	31
	50	20		50	38
0.15	10	8	0.30	10	12
	20	13		20	21
	30	18		30	28
	40	22		40	36
	50	26		50	43

TABLE II Behavior of $\bar{k}$ with respect to r and N for several values of r and N. Values of $\bar{k}$ computed with truncation error of 10^{-3} in $\emptyset^*$.

N	$\bar{k}$	$R^*(N,\bar{k})/T(N)$	Predicted time saving ratio	Actual running time (seconds)	Error in third significant figure
30	2	0.0002	0.0003	3.93	460
	3	0.0004	0.0006	6.64	286
	4	0.0010	0.0011	12.68	149
	5	0.0021	0.0019	22.02	66
	6	0.0040	0.0033	37.32	25
	7	0.0072	0.0053	60.10	8
	8	0.0121	0.0084	96.11	2
	9	0.0195	0.0123	140.24	Converged
	60		1.0	3.18 hrs. (estimated)	$\phi^* = 0.688$

TABLE III Computation time estimates for $\phi^*(29,0,1)$, with N=30, r=0.05, for several $\bar{k}$.

n_1 \ n_3	0	1	2	3	4	5	6	7	8	9	10
0	.050	.052	.054	.055	.057	.060	.064	.069	.076	.085	.095
1	.050	.053	.055	.058	.062	.067	.073	.081	.091	.104	
2	.049	.053	.058	.063	.069	.076	.087	.099	.114		
3	.049	.055	.062	.070	.079	.092	.108	.127			
4	.048	.058	.069	.082	.098	.118	.143				
5	.047	.063	.081	.104	.131	.163					
6	.047	.074	.105	.145	.190						
7	.046	.097	.158	.229							
8	.044	.158	.289								
9	.040	.397									
10	.016										

TABLE IV The error matrix $E^{(10)}$ for $\phi = |f(\underline{n}) - f(\underline{m})|$, r=0.05 and truncation error 10^{-3}.

n_1	$\phi^*(n_1,0,n_3)$	n_1	$\phi^*(n_1,0,n_3)$
9	0.6665	4	0.6661
8	0.6663	3	0.6660
7	0.6662	2	0.6659
6	0.6662	1	0.6657
5	0.6661	0	0.6655

Exact value (by Theorem 7.4) of $\phi^*(n_1,0,n_3)$ is $2/3 + 10^{-6}$

TABLE V Computed values of $\phi^*(n_1,0,n_3)$ for t=1, n=10 and $\delta=10^{-3}$.

k	Actual R_k	$\bar{R}_k$
1	9.50	10.5
2	6.00	6.67
3	4.24	4.76
4	3.21	3.60
5	2.50	2.84
6	2.00	2.29
7	1.63	1.88
8	1.33	1.56
9	1.10	1.30

TABLE VI R_k and estimates for N=10 and $k \leq 10$.

r	t	N	$\bar{k}$	ε/δ	$t.\bar{R}_{\bar{k}}$
0.05	0.053	20	11	0.166	0.14
		50	26	0.180	0.15
0.20	0.25	20	12	1.42	0.58
		50	33	1.11	0.51
0.30	0.43	20	21	0.75*	
		50	44	2.29	0.67
0.40	0.67	20	25	2*	
		40	45	2*	

TABLE VII Improved truncation error estimate.

*since $\bar{k} > N$, the second form of the bound for ε applies.