Feature Selection Approach for Diagnosis in Microbiology and Its Application to Yeast Species

LUTZ BRÜGGMANN,* INGEBORG FLEMMING, AND BERTRAM NAGEL
Institute of Biotechnology, Permoserstrasse 15, Leipzig 0-7050, Germany

By using mathematical logic and the fuzzy technique, we present an approach to the computation of optimal feature sets for diagnostic tests in microbiology. These sets of physiological and biochemical tests are suitable for differentiating all organisms belonging to a given group. Membership functions are defined which are able to reflect the time dependence of the test reaction states. The result is a suboptimal solution characterized by short feature sets of high diagnostic value. An application to Rhodotorula species is shown.

Classification and diagnostics in microbiology can be performed by means of qualitative feature patterns such as morphological, physiological, biochemical, and sometimes ecological characteristics. These characteristics can be used for differentiating between both genera and species. The solution of a diagnostic problem on the basis of given feature patterns leads to a need for dimension reduction, because good differentiation between individuals is achieved if a subset of features having a high diagnostic value is used (3, 8).

Generally, in pattern recognition the problem of selecting features with high separation power (as a mathematical optimization problem) is solved principally as follows (5): (i) Starting from a given feature set, subsets can be formed which minimize a given goodness criterion by means of search strategies, resulting in a suboptimal solution. (ii) Transformation techniques such as the Karhunen-Loeve transformation (problem-dependent optimal solution) or the Fourier transformation (general method) can be applied. These transformation techniques produce new artificial features with suitable properties (low dimension, orthogonality, no correlations).

For microbiological features with qualitative categories, only the first type of feature selection can be taken into account. This is important because it decreases the expenses for microbiological tests, rather than the expenses for computerized classification procedures. For yeasts, from a set of about 80 commonly used physiological and biochemical tests, each of which results in one of six qualitative categories (1), features with a high diagnostic value for separating a given group of yeast species should be selected. Therefore, optimal feature sets are calculated, whose use leads to minimized expenses. A further reason for the use of optimized feature sets is that the classical approach, in which hierarchical (dichotomous) determinative keys are used, is complicated if additional states (D for delayed response, S for slow utilization, W for weak utilization) as variations of the category + (positive response) are included. Identification keys based on the questionnaire theory (6, 7) provide polychotomous questionnaires with a complex structure. Furthermore, working in parallel with a series of tests based on feature sets is more efficient than working sequentially with hierarchical keys.

THEORY

Fundamentals. The basic idea for the determination of nonredundant feature sets was presented by Kautz (4). A logical expression is constructed for a given organism-test matrix. For a group of \( n \) organisms, a total of \( n(n - 1)/2 \) paired comparisons are performed, and from each compari-

* Corresponding author.

FIG. 1. Time dependence of membership functions referring to feature categories of physiological and biochemical tests.
son a disjunction of all possible tests is obtained which is able to distinguish the pair. The conjunction of these disjunctions gives a logical expression, \( H \). This expression is true if each comparison provides at least one separating test or, in other words, if the group of organisms is separable. \( H \) is a conjunctive normal form and can be transformed by means of Boolean algebra into an equivalent disjunctive normal form, \( N \), consisting of prime conjunctions. Thus, each prime conjunction of \( N \) refers to a minimal nonredundant feature set that is able to separate the group of organisms.

Working with the additional categories \( S, D, \) and \( W \) as fuzzy variations of + leads to an improvement in information. Even in this case the feature selection problem can be solved with respect to the real separation power of these categories. For instance, the separation power of a test with the response \( W \) is smaller than the separation power of a test with the response + or –. The purpose of modeling is to distinguish between organisms, including fuzzy feature categories. The fuzzy technique is an appropriate method for solving this problem. The time dependence of the test reaction is not interpreted as a stochastic process. Therefore, a statistical approach (see, e.g., reference 9) is less suitable. For these reasons the conventional approach based on mathematical logic was extended by using the fuzzy technique.

The fuzzy methods, which are based on the work of Zadeh (10), can be usefully applied to problems that cannot be considered as deterministic or stochastic. A fuzzy set \( A \) of \( X = \{x\} \) is characterized by a membership function \( f_x: X \rightarrow [0,1] \). The value \( f_x(x) \) represents the degree of membership for each point \( x \), in the fuzzy set \( A \) (Fig. 1). The membership function \( f_x \) can be defined objectively by probability distributions, but it can also be used to incorporate personal subjective experience. If \( f_x \) has only two values, 0 or 1, it is reduced to the conventional set theory, which is \( f_x(x) = 1 \) for \( x \in A \) and \( f_x(x) = 0 \) for \( x \notin A \).

Thus, each organism is considered a vector of membership functions of fuzzy sets, wherein these fuzzy sets represent the time-dependent response of corresponding features. For instance, the organisms \( O_k \) and \( O_l \) may be characterized by a \( p \)-dimensional vector in the following way:

\[
O_k = (\psi_{1,k}, \ldots, \psi_{p,k}), \quad O_l = (\psi_{1,l}, \ldots, \psi_{p,l}),
\]

where \( \psi_{ij} \) is the membership function of feature \( j \) with respect to organism \( i \).

A distance measure, \( \delta_j \), between \( O_k \) and \( O_l \) with respect to feature \( j \) may be defined as

\[
\delta_j(O_k, O_l) = \int_0^1 (\psi_{k,j}(t) - \psi_{l,j}(t)) dt
\]

if the time is normed on the interval \([0, 1]\).

It can be shown readily that \( \delta_j \) has the properties of a metric. The functions \( \psi_{ij} \) represent the five categories +, S, D, W, and –. When identified with the functions \( \varphi_1, \ldots, \varphi_5 \), they have the shape as shown in Fig. 1. These functions were arbitrarily chosen, but they should be simple algebraic functions and should, from the point of view of microbiologists, reflect the time and the state of the test reactions.

Now it is possible a priori to calculate the distance matrix \( D = (d_{rs}) \), where

\[
d_{rs} = \int_0^1 (\varphi_r(t) - \varphi_s(t)) dt \quad \text{with} \quad r,s \in \{1, \ldots, 5\}
\]

and

\[
D = \begin{bmatrix}
0 \\
0.25 & 0 \\
0.50 & 0.25 & 0 \\
0.78 & 0.54 & 0.44 & 0 \\
1.00 & 0.75 & 0.50 & 0.22 & 0
\end{bmatrix}
\]

\* Rhodotorula species and tests taken from reference 1.
\* Responses: D, delayed; W, weak; V, variable; ?, unknown.

| TABLE 1. Summary of 79 assimilation tests and other growth characteristics of Rhodotorula species* |
| Species no. | Species | Response to test no.* |
| 383 | R. acheniorum | + V V + + + + + + + + V V V + + |
| 384 | R. aracariarum | - - D - + + - + + + + + + + |
| 385 | R. aurantiaca | V V V V V V V V V V - D - - - |
| 386 | R. bogoriensis | - - - V V V V V V - - - D + + - + |
| 387 | R. diffusus | - - - - - - + + + + + + + + + + |
| 388 | R. gigas | V V V V V V V V V V V V V V V V |
| 389 | R. gilatinosa | + + + + + + + + + + + + + + + + |
| 390 | R. graminis | + + + + + + + + + + + + + + + + |
| 391 | R. ingeniosa | + + + + + + + + + + + + + + + + |
| 392 | R. javanica | + + + + + + + + + + + + + + + + |
| 393 | R. lactosa | - - - + + + + + + + + + + + + + |
| 394 | R. marina | + + + + D - - - + + + + + + + + |
| 395 | R. minata | + + + + + + + + + + + + + + + + |
| 396 | R. muscarum | - - - - - - - - - - + + + + + + |
| 397 | R. muscorum | + + + + + + + + + + + + + + + + |
| 398 | R. pallida | + + + + + + + + + + + + + + + + |
| 399 | R. pilattii | + + + + + + + + + + + + + + + + |
| 400 | R. pilimanae | + + + + + + + + + + + + + + + + |
As an example, in matrix $D$ the distance $d_{1,2}$ between the functions $\varphi_1$ (category $+$) and $\varphi_2$ (category $S$) is calculated as follows:

$$d_{1,2} = \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(t^2)} dt$$

$$= \int \left[1 - \left[(1 - \frac{1}{2})^2\right]\right] dt = 0.25$$

Thus, to each pair of organisms $(O_x, O_y)$ there belongs the following $p$-dimensional vector:

$$\delta(O_x, O_y) = \{\delta_1(O_x, O_y), \ldots, \delta_p(O_x, O_y)\}$$

TABLE 2. Minimal feature sets for the identification of Rhodotorula species in which the fuzzy character states D and W were treated by the fuzzy technique without optimization of separation power.

Program DPMINSET: Computation of minimal feature sets

Genus: Rhodotorula
Species: 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400
(cf. Table 1)

No. of pairs: 153
No. of sets: 12

Minimal feature set

<table>
<thead>
<tr>
<th>No. of pairs: 153</th>
<th>No. of sets: 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 22 29 32 48 49 54 72</td>
<td>16 22 29 32 48 49 54 72</td>
</tr>
<tr>
<td>16 22 29 32 48 49 54 72</td>
<td>16 22 29 32 48 49 54 72</td>
</tr>
<tr>
<td>16 22 29 32 48 49 54 72</td>
<td>16 22 29 32 48 49 54 72</td>
</tr>
<tr>
<td>16 22 29 32 48 49 54 72</td>
<td>16 22 29 32 48 49 54 72</td>
</tr>
<tr>
<td>16 22 29 32 48 49 54 72</td>
<td>16 22 29 32 48 49 54 72</td>
</tr>
<tr>
<td>16 22 29 32 48 49 54 72</td>
<td>16 22 29 32 48 49 54 72</td>
</tr>
<tr>
<td>16 22 29 32 48 49 54 72</td>
<td>16 22 29 32 48 49 54 72</td>
</tr>
<tr>
<td>16 22 29 32 48 49 54 72</td>
<td>16 22 29 32 48 49 54 72</td>
</tr>
<tr>
<td>16 22 29 32 48 49 54 72</td>
<td>16 22 29 32 48 49 54 72</td>
</tr>
<tr>
<td>16 22 29 32 48 49 54 72</td>
<td>16 22 29 32 48 49 54 72</td>
</tr>
<tr>
<td>16 22 29 32 48 49 54 72</td>
<td>16 22 29 32 48 49 54 72</td>
</tr>
<tr>
<td>16 22 29 32 48 49 54 72</td>
<td>16 22 29 32 48 49 54 72</td>
</tr>
<tr>
<td>16 22 29 32 48 49 54 72</td>
<td>16 22 29 32 48 49 54 72</td>
</tr>
</tbody>
</table>
it is now possible to determine feature sets. As a measure of the
diagnostic value of a feature set \( i \), a so-called mean separation power, \( s_i \), is defined:

\[
s_i = \left(1/n_i\right) \sum_{j=1}^{n_i} \delta_{ij}^{(s)}
\]

where \( \delta_{ij}^{(s)} \) is the resulting distance of feature \( j \) in set \( i \) referring to all \( n(n - 1)/2 \) species included and \( n_i \) is the number of features belonging to feature set \( i \). In the following discussion, an example will be treated both in the conventional
way and by the generalization given above. Let a species-test matrix be given with \( n = 3 \) species as organisms and \( p = 4 \) microbiological tests (a, b, c, and d) as features: In the conventional way, we have the distance vectors \( \delta(O_1, O_2) = (1,1,1,1) \), and \( \delta(O_2, O_3) = (1,0,1,1) \), and the minimal disjunctive normal form is calculated by means of Boolean algebra:

\[
\begin{align*}
H_1 &= (b + c + d) \cdot (a + b + c + d) \cdot (a + c + d) \\
&= (ab + bb + bc + bd + ac + bc + cc + cd + ad + bd + cd + dd) \cdot (a + c + d)
\end{align*}
\]

\[
H_2 = (b_0.5 + c_0.25 + d_1) \cdot (a_1 + b_0.5 + c_1 + d_0.75) \cdot (a_1 + c_0.75 + d_0.25)
\]

\[
= (a_1b_0.5 + b_0.5 + b_0.5c_1 + b_0.5d_0.78 + a_1c_0.25 + b_0.5c_0.25 + c_0.25 + c_0.25d_0.38 + a_1d_1 + b_0.5d_1 + c_1d_1 + d_0.78) \cdot (a_1 + c_0.75 + d_0.25)
\]

\[
= (b_0.5 + c_0.25 + d_0.78 + a_1d_1 + c_1d_1) \cdot (a_1 + c_0.75 + d_0.25)
\]

\[
= a_1b_0.5 + b_0.5c_0.25 + b_0.5d_0.22 + a_1c_0.25 + c_1d_1 + c_1d_1 + d_0.78 + c_0.25d_0.28 + a_1d_1 + a_1d_1 + a_1d_1 + a_1d_1 + a_1c_0.25 + c_0.25d_0.28 + c_1d_1 + c_1d_1 + d_0.78
\]

\[
= a_1b_0.5 + b_0.5 + c_0.25 + d_0.25 + d_0.22 \quad \text{or} \quad N_2
\]
Obviously, in this case the second feature set, \(a_1d_1\), has the largest diagnostic value (\(s_2 = 1\)) of all six possible feature sets.

**Computation method.** The disjunctive normal form described above is constructed step by step by means of fuzzy algebra. The number of feature sets increases with a factor \(m\) if \(m\) new separating features were found at the next paired comparison. The length of expressions is reduced by rules of shortening. Nevertheless, the increase in the number of sets leads to a data explosion and to an unrealistic computation time when a larger number of organisms are included. Therefore, during the computation the number of feature sets is limited by the available internal storage capacity. Only the feature sets with a large value of \(s_i\) are used in the next step. As a result, a suboptimal solution is achieved. The data are processed by access to a feature data file of the Biotechnology Data Bank BD1 (2).

**RESULTS AND DISCUSSION**

Table 1 lists the character states of 79 assimilation tests and other growth characteristics of 18 *Rhodotorula* species (taken from reference 1; some tests were modified) which have been used as initial data. Only two fuzzy categories, \(D\) and \(W\), occur within these character states.

Table 2 shows minimal feature sets which were calculated on the basis of conventional logical operations. The results shown in Table 3 were obtained by applying fuzzy rules; they differ slightly from those in Table 2. From these results, the real diagnostic value of the feature sets can be seen by the separation power \(s_i\) (0 < \(s_i\) ≤ 1; \(s_i = 1\) if essentially only the categories + and − are used). Table 4 shows minimal feature sets as optimized for a maximal value of \(s_i\). For only one set (\(s_{15} = 0.95\)), one fuzzy test is essential for diagnosis. The higher quality of the feature sets of Table 4 can be seen by comparing the first feature set (first row) of Tables 2 and 4 with respect to the occurrence of fuzzy feature states in the respective columns:

The allocation matrix in Table 5 was compiled on the basis of the first feature set of Table 4 as a diagnostic key for the *Rhodotorula* species. As seen by comparison with the initial data (Table 1), the decreasing expenses needed for the respective diagnostic problem are evident.

If a diagnostic problem must be solved within an extremely short period, the approach presented here can be slightly modified to deliver test sets which have a high separation power within this time frame.

**ACKNOWLEDGMENT**

We thank K.-U. Jahn, Technical University Leipzig, for helpful discussions.

**REFERENCES**


