Developing Classification Techniques from Biological Databases using Simulated Annealing

B. de la Iglesia (bli@sys.uea.ac.uk), J.J.Wesselink (jjw@sys.uea.ac.uk) and V. J. Rayward-Smith (vjrs@sys.uea.ac.uk)
School of Information Systems, University of East Anglia,Norwich, England

J. Dicks (jo.dicks@bbsrc.ac.uk)
John Innes Centre, Norwich Research Park, Colney,Norwich, England

I. N. Robert (ian.roberts@bbsrc.ac.uk)
Institute of Food Research, Norwich Research Park, Colney, Norwich, England

V. Roberts (robert@cbs.knaw.nl) and T.Boekhout (boekhout@cbs.knaw.nl)
Centraalbureau voor Schimmelcultures,Utrecht, The Netherlands

Abstract. This paper describes new approaches to classification/identification of biological data. It is expected that the work may be extensible to other domains such as the medical domain or fault diagnostic problems.

Organisms are often classified according to the value of tests which are used for measuring some characteristic of the organism. When selecting a suitable test set it is important to choose one of minimum cost. Equally, when classification models are constructed for the posterior identification of unnamed individuals it is important to produce optimal models in terms of identification performance and cost.

In this paper, we first describe the problem of selecting an economic test set for classification. We develop a criterion for differentiation of organisms which may encompass fuzzy differentiability. Then, we describe the problem of using batches of tests sequentially for identification of unknown organisms, and we explore the problem of constructing the best sequence of batches of tests in terms of cost and identification performance. We discuss how metaheuristic algorithms may be used in the solution of these problems. We also present an application of the above to the problem of yeast classification and identification.

Keywords: classification, identification, Minimum Test Set (MTS), heuristic techniques

1. Introduction

Heuristic algorithms for mining large databases are being adapted to enable discriminatory analysis to be performed on biological data, accelerating the progress in understanding biological diversity and its industrial implications. A range of knowledge discovery algorithms are being applied to yeast characteristics data, providing new research leads and decision making tools. The research presented here is part of a project funded by the BBSRC [1] which involves the curation and data mining analysis of yeast species and strain data, including DNA data for 700+ yeast species. There is special industrial interest in the investigation of yeast species capable of causing food spoilage, including emerging spoilage food agents.

The initial phases of the project, which cover some of the work reported here, have focused on understanding and improving the current methods for identification and classification in the biological domain. The new methods being developed should lead to faster, cheaper and more reliable classification and identification. Yeast data is being used as an initial case study, but it is expected that the approach can be extended to other taxa in the biological domain, and possibly to other domains such as medical diagnosis.

The problem of identification/classification in the biological domain has received attention in the literature for decades (Pankhurst, 1975; Payne, 1991; Payne and Thompson, 1989; Willcox and Lapage, 1972). According to Payne (1993), the aim is to find a model for identifying taxa (i.e. species, genera, populations, etc.) whose properties can be defined in terms of tests each of which has only a finite number of possible results. The term test is used to provide a general terminology. For example, in botanical identification each test may take the form of determining which state of a particular character is exhibited by the specimen being identified (e.g. the test "colour of petals" may have values "red", "blue","yellow"). In yeast, a test may represent the ability of a specimen to use various compounds for growth (e.g. the test "D-Xylose Growth" may have results "positive", "negative" and "equivocal"). The problem of identification is not only confined to the biological domain, though. In electronic fault diagnosis, a test may involve seeing whether a particular set of components contains a fault, and in medical diagnosis, a test may contain recorded signs or symptoms for a patient.

Traditionally in biology, diagnostic(or identification) keys or diagnostic tables were constructed and used for identification, see for example (Payne and Preece, 1981; Payne, 1992). A diagnostic key is a type of decision tree, for those familiar with Knowledge Discovery in

Databases (KDD). A diagnostic table is simply a table of results listing the outcome of each test or experiment for each organism. To produce diagnostic keys or tables, it is sometimes necessary to find a minimum characteristic set (or test set) containing enough characteristics to discriminate all organisms. Payne (1991) describes an algorithm which can be used to find all irredundant test sets (i.e. those containing no tests which can be omitted without making any organisms unidentifiable). The Minimum Test Set (MTS) or the test set of minimum cost (if costs are given) capable of classifying all organisms can be chosen from the list of irredundant test sets, as test sets with redundant tests could not be minimum. However, Payne's algorithm has exponential worse case complexity. In fact, the problem of finding a Minimum Test Set (MTS) to differentiate organisms is NP-hard according to Garey and Johnson (1979). For tests with binary outcomes, finding the optimal decision tree is NP-hard (Hyafil and Rivest, 1976). It follows that, constructing a key with the minimum expected cost of identification is NP-hard. A heuristic approach to key construction, very similar to the approach used for constructing decision trees (Quinlan, 1993; Quinlan, 1996), has been used in practice. The approach is based on selecting sequentially a test that "best" divides the organisms into sets. Intuitively, the best test may be that which more nearly splits the taxa into groups of equal size. However, if equivocal or variable responses to tests are present the choice is not so simple. Various selection criteria for the tests have been used and evaluated (Payne, 1981; Payne and Dixon, 1984; Payne and Preece, 1981; Payne and Thompson, 1989). Programs such as Genkey (Payne, 1993) can be used to produce diagnostic tables and diagnostic keys.

The problem of choosing a classification or identification model is closely related to that of choosing a classification model in data mining. There are also some parallels with Feature Subset Selection (FSS) (Debuse and Rayward-Smith, 1997; Liu et al., 1998; Kononenko, 1994; Hall, 1998), which is another tasks of the pre-processing stage in the KDD process (Debuse et al., 2000). The objective of FSS is to isolate a small subset of highly discriminant features and this has some parallels with finding a MTS.

This paper will first report how it may be possible to use metaheuristics, such as Simulated Annealing previously used for FSS and data mining (de la Iglesia et al., 1996; de la Iglesia and Rayward-Smith, 2001), to the problem of finding a MTS. To do this, we first model the problem as a combinatorial optimisation problem. The modeling involves establishing a method that allows organisms to be distinguished by some fuzzy criteria, when a crisp differentiation is not possible. The objective of the metaheuristic algorithm will be to choose the classifi-

cation model which allows for maximum differentiability of organisms, while minimising the total number of characteristics (or the total cost of characteristics, if this is given) required for classification. The problem of delivering a model that will allow for "economic" identification of unnamed organisms using batches of tests will also be discussed. For identification, the objective is to choose a model which will maximise the chances of delivering an identity for an unknown organism while minimising the total cost of the identification process.

Organisms may be assigned a weight, so that if it is crucial to identify a particular organism within the model induced, the weights can be adjusted to give this a priority. The weight may for example be used to indicate known frequency of occurrence, in which case the model will represent minimising expected cost of classification/identification. Alternatively, a relatively rare organism, but one that is of particular importance, may be given a high weight to ensure its identification. The approach has immediate application to the determination of the tests necessary to identify types of yeast, such as food spoilage agents, which are of great industrial interest.

2. Differentiability of organism

The data will consist of a number of records which contain a series of characteristics with their values for different organisms. Characteristics recorded may have an associated cost value. Organisms may have a weight. It is worth noting that with the available data it may not be possible to obtain a perfect classifier (i.e. one that is capable of distinguishing all available organisms); also some characteristics may be redundant to the classification. In this context, the definition of an optimal model may vary but will incorporate the idea of "economy" of classification/identification.

More formally, let us consider a set of characteristics

$$P = \{t_1, t_2, \ldots, t_n\}$$

which applies to each organism $s \in S$ (S is the set of all known organisms to be classified). Each t_i may have an associated cost value $c(t_i) \in \mathbb{R}^+$. Each organism may have a weight, $w(s) \in]0, 1[$ which may, for example, account for the frequency with which they occur. Each organism is initially defined by a set of values for each characteristic in P; the value of characteristic i for organism s is given by $s[t_i] \in \mathrm{Dom}_i$.

We define an organism s as *differentiable* if, for a given set of characteristics, $Q \subseteq P$, the combination of values for s, $s[Q]$, is unique in S, i.e. $s[Q] = s'[Q] \Rightarrow s = s'$.

In some applications, uniqueness is not strong enough for differentiability since some values, although distinct are not significantly different, e.g. yeast experts consider a <positive> and a <positive, weak, delayed> response to a growth test as equivocal. Hence, in some cases the concept of differentiability is *fuzzy*. To store the perceived difference between two test responses, we use a differentiability function. This function records the difference value assigned to each pair of characteristic values

$$diff_i \in \mathrm{Dom}_i \times \mathrm{Dom}_i \rightarrow [0,1].$$

If $diff_i$ only takes a value of 0 or 1, then we have *crisp* differentiability of organisms, otherwise we have *fuzzy* differentiability. *diff* will denote the n-tuple of differences

$$(diff_1, diff_2, \ldots, diff_n).$$

Associated with any set of characteristics, $Q \subseteq P$, any *diff*, and any $s \in S$, is a measure $\delta(s, Q, diff)$ where $\delta(s, Q, diff)$ is equal to zero if s is not differentiable from all other species in S according to tests in Q and it is one otherwise. To compute the *total differentiability* of an organism, $\delta(s, Q, diff)$, we must first compute the differentiability between that organism and each of the other organisms. The diffentiability between two organisms $s \in S$ and $s' \in S$ in $Q = \{t_{\lambda 1}, t_{\lambda 2}, \ldots, t_{\lambda n}\}$ is computed as

$$\begin{aligned}\delta(s, s', Q, diff) = \quad & diff_{\lambda 1}(s[t_{\lambda 1}], s'[t_{\lambda 1}]) \\ & * \ diff_{\lambda 2}(s[t_{\lambda 2}], s'[t_{\lambda 2}]) \\ & * \ \ldots \ * diff_{\lambda n}(s[t_{\lambda n}], s'[t_{\lambda n}]).\end{aligned}$$

The operator $*$ needs to satisfy certain properties. For any $x, y, z \in [0,1]$, $*$ is an operator $[0,1]^2 \rightarrow [0,1]$ that must satisfy

$$\begin{aligned}& x * y = y * x && \text{(commutative)} \\ & x * (y * z) = (x * y) * z && \text{(associative)} \\ & x_1 \leq x_2 \text{ implies } x_1 * y \leq x_2 * y && \text{(non-decreasing)} \\ & x * 1 = 1; x * 0 = x \\ & x * y \geq x\end{aligned}$$

i.e. $*$ is a t-conorm (Zimmermann, 1991). One example of a t-conorm which could be used for $*$ is the max-operator. Another possibility is to define $*$ by

$$x * y = x + y - xy.$$

Once the differentiability between each pair of organisms,

$$\delta(s, s_i, Q, diff) : \forall s_i \in S, s_i \neq s$$

is calculated then the total differentiability of an organisms, $\delta(s, Q, \mathit{diff})$ is obtained as

$$\delta(s, Q, \mathit{diff}) = 1, \text{ if} \min\{\delta(s, s_i, Q, \mathit{diff}) : s_i \in S,\ s_i \neq s\} \geq M$$
$$\delta(s, Q, \mathit{diff}) = 0, \text{ otherwise,}$$

where M is a threshold parameter which can be adjusted to allow different degrees of differentiability. According to the previous definitions of $*$, if M is set to 1, then the criterion for differentiability will be non-fuzzy, whereas if M is set to a value less than 1, then the criterion will be fuzzy.

If $\delta(s, Q, \mathit{diff})$ is equal to 1 we say that organism s is uniquely identifiable in S according to the tests in Q and to the differentiability matrix *diff*.

3. Choosing a test set for classification

The previous calculations allow us to compute the differentiability of any organism with respect to the rest of organisms in S, according to any subset of characteristics, $Q \subseteq P$, and using a differentiability function, *diff*, which can encompass fuzzy or crisp differentiability criteria.

If the problem to be solved is to find the Minimum Test Set (MTS) capable of classifying the organisms of maximum combined weight (equivalent to the maximum number of organisms when weights are not used), then we must choose the subset of characteristics $Q \subseteq P$ that minimises

$$\sum_{t \in Q} c(t),$$

subject to the following constraint

$$\sum_{\forall s \in S} \delta(s, Q, \mathit{diff}) \times w(s) = \sum_{\forall s \in S} \delta(s, P, \mathit{diff}) \times w(s).$$

However, we may be able to approach this by solving a wider problem. It may be possible, and perhaps desirable in some situations, to trade-off the unique identification of some organisms for a lower cost of the test set Q. If the criterion of cost is set against the criterion of performance of the model, in terms of the number of identifications achievable, then the problem becomes a pareto optimisation or multi-optimisation problem (Fonseca and Fleming, 1995; Osyczka, 1985). We then want to choose the subset of characteristics, $Q \subseteq P$, that maximises

$$\sum_{\forall s \in S} \delta(s, Q, \mathit{diff}) \times w(s),$$

and minimises

$$\sum_{t \in Q} c(t).$$

This problem will have a number of non-dominated solutions, in pareto optimisation terms, which form the set of pareto optimal solutions.

Modern heuristics are widely used in pareto optimisation (Horn and Nafpliotis, 1994; Parks and Miller, 1998; Srinivas and Deb, 1994). Often in multi-objective problems, objectives are artificially combined into a scalar function, for example by the use of simple weighted sums (Jakob et al., 1992). Other approaches treat objectives separately, for example Schaffer (1985) uses different objectives to select sub-populations in a genetic algorithm and then merges and shuffles all sub-populations continuing the process of mating in a normal way but monitoring the population for non-dominated solutions. Approaches using Pareto-based genetic algorithms (Goldberg, 1989) assign equal probability of reproduction to all non-dominated individuals in the population. Recent developments have outperformed previous approaches. For example, the pareto-based algorithm NSGA II (Deb et al., 2000) (based on the sorting of solutions according to the concept of non-domination) has been shown to be efficient and effective at finding good approximations to the pareto optimal front.

An alternative approach to performing pareto optimisation may be to use a bound on the total cost of the characteristics contained in Q, i.e.

$$\sum_{t \in Q} c(t) \leq B,$$

where B is a cost bound. If we start, for example, with a high value of B and repeat the search for decreasing values of B then we may be able to sample solutions from the pareto optimal set using a simple meta-heuristic. In fact, a single solution obtained in this way may be all that is required in some practical applications. This is true of many pareto optimality problems in which a single compromise solution is usually sought. We can perform this optimisation using simulated annealing, and we present a possible implementation in section 5.

Whatever way the problem is solved, the resulting test set for the organisms identified could be used as a diagnostic table, which can be used for identification. When trying to identify a new organism, the test set values would be checked to try to find one row of test results matching the values of the organism being identified. A test set of minimum cost can also be used to construct an identification model.

4. Selecting a model for identification using batches of tests

We denote by identification the process of finding the classification of an unknown organism. Note that, in yeast identification for example, it is common practice to perform all tests necessary for classification, i.e. all the tests that form part of the MTS, in advance of the identification process. However, it is possible that identification of some organisms can be delivered by doing a batch of the tests in Q, with the rest of the tests performed if the results of the first batch do not produce an identification. This may be particularly suitable if some tests are difficult to perform simultaneously. The first batch could be constrained to contain tests that can be performed simultaneously in an economic way, with other tests being done if the results of the first batch prove inconclusive. As a hypothetical example, to identify a food spoilage yeasts it may be possible to perform only some fermentation tests which may be cheaply performed simultaneously.

4.1. Problem formulation for tests without equivocal responses

Any set of characteristics, Q, partitions the organisms into classes or groups of organisms indistinguishable by Q. In the model described in the previous section, the ideal is to produce a cheap set where the associated classes each contain at most one organism as this would deliver a perfect classification. Now, we consider a sequence of tests.

Instead of selecting all tests, the new approach involves selecting an initial batch of experiments $B \subseteq Q$. The initial batch would produce some classes of organisms, some with more than one element, and those could be further partitioned by an additional set of tests in the next stage, repeating this process until a unique organism is identified, or the tests are exhausted. This approach is applicable when there are no equivocal tests values in the database (i.e. no unknown, variable, or other equivocal results are recorded for any tests). As long as tests are unequivocal, then classes form a partition of S. However, if there are tests with equivocal outcomes, then organisms may appear in more than one class and therefore S cannot be partitioned into disjoined groups.

Let T be the tree induced by some experimental procedure. Each internal node in T represents a batch of experiments, and each leaf node represents a partitioning of S. If a leaf node comprises a singleton set, $\{s\}$, then we set $\delta(s,T) = 1$, otherwise $\delta(s,T) = 0$. We denote by cost(s) the cost of all the experiments performed at each node on the path from the root to the leaf containing s. For identification, we seek

a tree, T, such that

$$\sum_{s \in S} \delta(s, T) * w(s)$$

is maximised subject to

$$\sum_{s \in S} \{\text{cost}(s) : \delta(s, T) \neq 0\} \times w(s) \leq B$$

for some bound B.

We have modeled the batch problem only for unequivocal tests. Let us examine the procedure for constructing trees. Since branches represent combinations of tests values and there are k possible outcomes to each test, if we choose m tests to form part of a batch, B, there are potentially m^k combinations of test values. An approach which tried to construct branches of the tree by examining each possible combination of test values would not be practical for large values of m and k. Naturally, some of the combination of test values may not exist in the real data.

Instead of examining all branches, we could examine only those combinations of test values that have support in the data. To do this we could perform the following procedure. First we would start by comparing each organism with every other organism to decide if each pair are differentiable according to the tests in batch B. Pairs of organisms that cannot be differentiated need to be placed in the same group, i.e. they need to follow the same branch in the tree. This is fairly straightforward procedure for tests with unequivocal responses.

4.2. Problem formulation for tests with equivocal responses

When we have to cater for equivocal outcomes, and this is realistically the case for most biological applications, the previous approach cannot be used. Suppose that we are going to build a tree, as presented above, for a set of tests with equivocal responses. When some of the tests considered have equivocal outcomes an organism may follow multiple branches in the tree, and therefore appear in a number of classes of indistinguishable organisms. For example, suppose that we have three organisms s_i, s_j and s_k with the values shown on table I for tests t_1 to t_3. In this case, s_i cannot be distinguished from s_j or from s_k according to the batch of tests $B = \{t_1, t_2, t_3\}$, but s_j can be distinguished from s_k. In this context,therefore, $s_i[B] = s_j[B]$ and $s_j[B] = s_k[B]$ does not imply $s_i[B] = s_k[B]$. Hence, if we group those organisms according to the results of the batch B, there will be two separate groups, one containing s_i and s_j, and another one containing s_i and s_k. Note that

Table I. A simple example with variable or unknown results

Organism	t_1	t_2	t_3
s_i	+	+	v
s_j	+	v	-
s_k	v	+	+

s_j and s_k should not be placed in the same group. If at the next stage of tree building, the node containing s_i and s_j is partitioned by another batch of tests, then organism s_i may appear as distinguishable. However, if the node containing s_i and s_k cannot be further partitioned by any batch of tests, then s_i is not identifiable in the tree. The previous modelling approach could not be used in this circumstances.

Furthermore, at the end of the initial comparison of pairs of organisms, the output is a list of pairs of organisms that are undistiguishable according to the tests in B. These pairs need to be considered for merging with other pairs of indistinguishable organisms to form the final groups or classes. For tests with unequivocal responses, this is fairly straightforward because in this case it would follow that if $s_i[B] = s_j[B]$ and $s_i[B] = s_k[B]$ then $s_j[B] = s_k[B]$, and hence s_i, s_j and s_k must be placed together in a node. In the case of organisms with equivocal responses,however, the problem is much harder to solve. If we represent the organisms as nodes in an undirected graph, and the pairs as edges joining the nodes then we can convert this into a graph problem. For example, suppose there is a set of organisms $S = \{s_1, \cdots, s_9\}$ which produces the following list of pairs of indistinguishable organisms for some batch of tests B with equivocal responses:

$$(s_1, s_6), (s_2, s_4), (s_5, s_8), (s_5, s_9), (s_6, s_9), (s_8, s_9).$$

The corresponding graph appears in figure 1. We note that each maximal clique in the graph (i.e. each clique that is not a proper subset of any other clique) represents one of the final groupings of indistinguishable organisms:

- Each unconnected organism (a clique of size 1) represents a distinguishable organism.
- Two nodes connected to one another but with no other common neighbour (a clique of size 2) represent a group with two organisms, for example s_6 cannot be distinguished from s_1 or s_9, but s_1 can

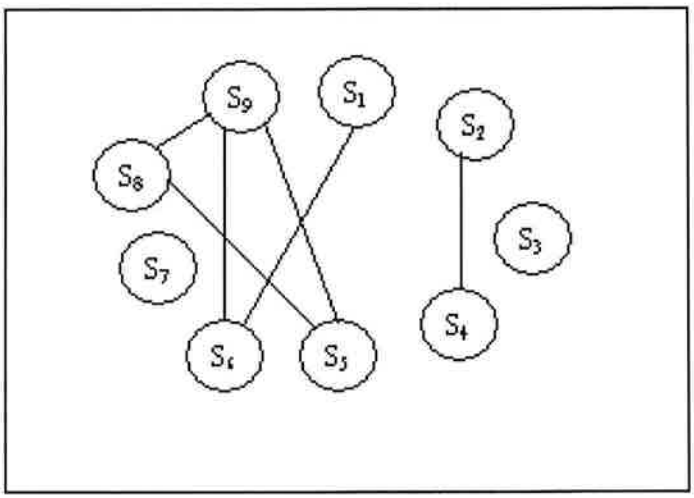

Figure 1. An undirected graph used to represent indistinguishable organisms

be distinguished from s_9 hence s_1 and s_9 must form a group on their own.

– s_5, s_8 and s_9 can be merged together as none of them can be distinguished from the others, and in the graph they form a maximal clique of size 3.

Hence the problem of finding how organisms are grouped into classes of indistinguishable organisms according to batch B when the batch contains equivocal tests can be solved by finding all the maximal cliques in a graph in which each node represents an organism and each edge represents a pairing of indistinguishable organisms. Unfortunately, this could be a rather hard problem to solve! In the worst case, there is an exponential number of maximal cliques in a graph. To prove this, consider the complement of the graph comprising $n/2$ cliques of size 2. This has $2^{n/2}$ maximal cliques.

Finding the maximum clique (i.e. the maximal clique that has the maximum cardinality)is an NP-hard problem (Karp, 1972). The maximal clique problem is even harder, in fact intractable for non-trivial problems, as the solution size may grow exponentially. This may signal the fact that the problem is not defined realistically.

4.3. Alternative "tractable" formulation

For tests with equivocal outcomes, it is necessary to find a different problem formulation. We can constrain ourselves to finding an initial batch of experiments to differentiate between some organisms, and then finding a second batch which would differentiate between the remaining organisms. Of course, in doing this, we loose the ability to determine later batches of tests depending on the outcome of the first batch of tests. However, this is a more tractable problem. Let us have a solution consisting of a set of batches $B_1, B_2, \ldots, B_k$. The application of batch B_i will produce a partition R_i consisting of all unidentified organisms. In order to evaluate a solution of this type, all organisms will be compared according to tests included in the first batch B_1. Then all those not identified by the first batch will form R_1 and will be compared according to tests in batch B_2, etc. Note that the costs of the first batch $cost(B_1)$ has to be added to the identification of all those organisms identified by the second batch, etc. We denote by p_i the total weight of organisms (or total number of organisms if weights are not used) identifiable in partition R_i using batch B_i, i.e.

$$p_i = \sum_{s \in R_i} \delta(s, \bigcup_{k=1}^{i} B_k, diff) * w(s),$$

where $\delta(s, B_i, diff)$ is equal to 1 is s is identifiable in partition R_i according to tests in B_i and 0 otherwise. The total cost of the solution can be calculated as follows:

$$\text{Total cost}\{B_1, B_2, \ldots B_k\} = \sum_{i=1}^{k} \left(\sum_{j=1}^{i} \text{cost}(B_j) \right) \times p_i.$$

Similar to previous formulations, we need to maximise the total weight (or number) of identifiable organisms according to the set of batches, i.e. maximise

$$\sum_{\forall s \in S} \delta(s, \{B_1, \ldots, B_n\}, diff) \times w(s),$$

and minimises the total cost as calculated in the previous equation. We can again approach this pareto optimisation problem by setting a bound on the total cost of the solution. Alternatively, we may add a large cost penalty for each organism that is not identified with the subsequent batches of tests, and simply concentrate on minimising the total cost of the solution.

5. Simulated Annealing implementation

A general purpose Simulated Annealing toolkit, such as *SAmson* (Mann, 1996), can be used as the platform for the implementation of the solution to the problem of finding the MTS. SAmson is an easy to use environment for the development of optimisation problems using Simulated Annealing. The SAmson toolkit consists of an intuitive interface including binary, integer and floating point representations. Adjustable parameters include the initial temperature, various cooling schedules, and different conditions for temperature equilibria. Various neighbouhood operators are implemented, and user defined operators can easily be incorporated. There is also a combined, random and adaptive method for neighbourhood operator selection. The main code in SAmson is implemented in the C programming language.

Sampling of the pareto optimal set can also be achieved with SAmson, by use of a cost bound. Sampling of the pareto optimal set should also deliver an approximation to the optimal MTS.

The implementation for this problem can be approached as follows:

- A binary alphabet can be used to represent a solution. Each position represents a characteristic. A 1 in position i implies that characteristic i is included in the model; 0 implies that characteristic i is excluded. This establishes the set of characteristics Q
- A solution is initialised by randomly setting some bits to 1, and checking that the total cost, $\sum_{t \in Q} c(t)$ is less than or equal to a user defined bound B. If the total cost exceeds the bound the initial solution can be modified or reconstructed until a satisfactory initial solution is reached. Alternatively, a penalty can be applied at the time of evaluation, proportional to the cost of the solution. In the case where costs of characteristics are unknown, it is only necessary to set less than or equal to B bits in the initially solution.

5.1. Neighbourhood structures

A new solution is reached by including a new characteristic previously excluded, excluding a characteristic previously included, or both. For each new solution, it is necessary to check that the bound on the total cost of the set Q is not exceeded. If the bound is exceeded, two approaches can be used:

- In the first more restrictive approach, the neighbourhood operator would reconstruct any solution that exceeds the total cost bound,

so that any solution produced is guaranteed to be a valid solution in terms of the cost bound B. This may lead to the neighbourhood operator slowing down the algorithm.

- In the second approach, the neighbourhood operator would be allowed to produce invalid solutions, in terms of the bound B. The evaluation function would penalise invalid solutions to ensure that the search is encouraged in a different direction.

To evaluate a solution, the total number of differentiable organisms using the characteristic set Q is computed, using the differentiability criterion described in section 2. If penalties are to be applied to invalid solutions, a penalty which is related to the cost of the characteristic set Q, if this exceeds bound B, can be added at this stage.

Experiments to choose suitable SA parameters such as cooling schedule, initial temperature, etc, will have to be performed to establish adequate parameter values.

The previous implementation can be changed by modifying the solution representation. For this approach, a solution can be represented by an array of n integers, where n is the total number of tests. The values of the array will be in the range [0..Max. no of batches]. Note that a value of 0 would mean that a test is excluded from all batches, hence this approach will also solve the previous problem. For each position in the solution string the corresponding array value represents the first batch number in which the test is included.

The inialisation process will randomly set each position in the solution array to a value in the range [0..Max. no of batches].

The neighbourhood operator will change one of the positions in the solution array to a different value. For example, a test previously excluded, may not be include in batch 2.

The evaluation of a solution will be performed by looking at all organisms and computing the number (or combined weight) of organisms identified by the first batch; the process will continue by selecting the subset of organisms still unidentified, and computing the number (or combined weight) of organisms identified by the second batch, etc. At the end of this process the total cost of the solution will be calculated as described previously. A penalty may be added for each unidentified organism.

The algorithm would be set to minimise the total cost of the solution generated.

6. Computational experiments and results

Motivation for this work was the need to investigate the classification of yeasts and in particular the characteristics of food spoilage yeasts not shared by other yeasts.

There are 10 species which have been identified by the experts as food spoilers (Pitt and Hocking, 1997), hence yeasts associated with food spoilage represent a minority of species. Nevertheless, they may cause very significant losses to the food and beverages industry, and therefore the isolation of characteristics that distinguish them from other yeast species is an important exercise.

Some of the yeast data available was donated by the Centraalbureau voor Schimmelcultures (CBS) [2]. The CBS maintains the largest collection of living fungi in the world. Their yeast database contains data on more than 4,500 strains kept in the Netherlands, and on about 1,250 strains from the IGC collection in Oeiras, Portugal. They also have data at the species level, for 745 species.

The CBS data stored for each yeast, at species or at strain level, consists of what we will name as *conventional* characteristics used to classify the yeasts. There are 96 characteristics recorded for each yeast species and they include microscopical appearance of the cells; mode of sexual reproduction; certain physiological activities and certain biochemical features. Each characteristic is recorded as having a particular response . The responses can be "positive", "negative", "positive, weak, delayed","negative, weak, delayed", "weak", "delayed" and "other". Explanation of characteristics used for the classification of yeasts and the meaning of their responses is given in (Barnett et al., 2000).

When comparing characteristic values for yeast, only negative and positive responses to characteristics are considered traditionally as establishing a difference, all other responses are considered as equivocal (Barnett, 1971a; Barnett, 1971b). The differentiability function for the yeast is shown in table II. Note that this differentiability function is applicable to all characteristics recorded since all draw their values from the same domain. This *diff* function produces a "crisp" differentiation of species, i.e. minor differences are not considered as sufficient for differentiation. For example, the difference between a positive and a delayed response is not considered significant. Yet, with this differentiability function, we found that many species cannot be differentiated using conventional tests. In particular, none of the spoilage yeasts present in the database were differentiable according to this criterion. This motivated the need for "fuzzy" differentiability criterion, which would allow minor differences (for example, the difference between a positive or a delayed response to a test) to count partially towards the

Table II. Differentiability matrix for the yeast

Responses	−	+	others
-	0	1	0
+	1	0	0
others	0	0	0

Table III. Fuzzy differentiability matrix for the yeast

Responses	−	+	$+, w, d$	$-, w, d$	w	d	others
-		1	0.5	0	0.5	0.5	0
+			0.25	0.5	0.5	0.5	0
+,w,d				0.25	0	0	0
-,w,d					0	0	0
w						0.25	0
d							0

differentiation of two species or strains. A fuzzy classification model results in higher differentiability between species/strains. For example using the crisp *diff* function provided in table II, only 544 out of 745 yeast species are identifiable, whereas using the fuzzy *diff* function provided in table III 643 species are identifiable using a threshold parameter M of 0.5 for calculating the totall differentiability of an organism, and 687 are identifiable with a threshold $M = 0.25$. The t-conorm operator used to compute the differentiability between two organisms in these experiments is $x * y = x + y - xy$, as that gives better results than the max-operator. This may be due to the fact that the combination of two values using the first operator results in a higher value of the differentiability between two organisms. An adequate value of M is obviously specific to each particular application, and can only be established by experimentation and by interpretation of the results.

Traditional classification methods for yeast use the different responses to conventional characteristics tests as a criterion for differentiability of species/strains. At present, comparisons of genomes in terms of base sequences are increasingly used for classification. The creation of a classification system based entirely on DNA sequence data is being investigated separately, and some of the work is reported in (Wesselink et al., 2002). Combination of both approaches is expected in the future.

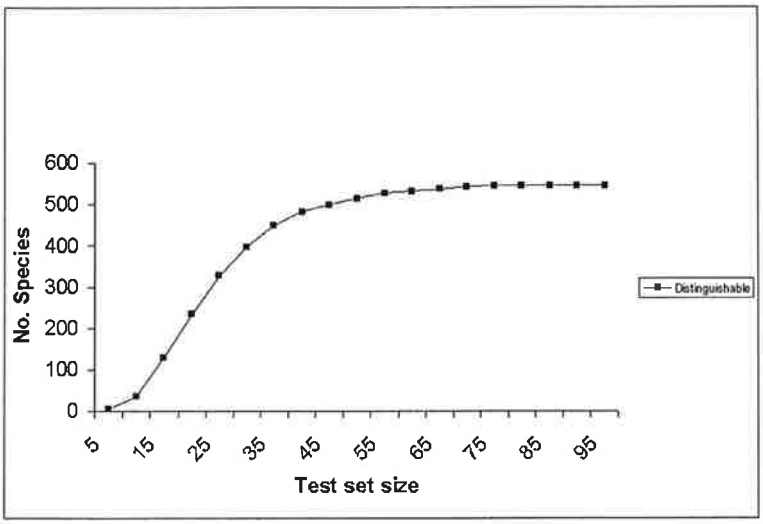

Figure 2. Some results for yeast data obtained with SA

Initial experiments with the CBS species data have highlighted sets of tests that are redundant to the classification model. This results were obtained using crisp differentiability, so they represent a lower bound. Figure 2 shows the number of differentiable species (as weights are assumed equal for this problem) for a single set of tests of varying sizes. The Minimum Test Set, according to this results, contains 75 tests. If minimising cost is of essence, a test set of 45 tests would achieve differentiability of most of the species in the dataset.

7. Conclusions and further research

In this paper we have discussed how to model classification and identification of organisms as combinatorial optimisation problems. The concept of fuzzy differentiability, introduced in our modelling, should result in higher differentiability of organisms.

We have shown how heuristic algorithms, and in particular Simulated Annealing, can be used to solve the formulated problem. We have also shown that trying to group organisms according to the results of a batch of tests may lead to an exponential partitioning step, and hence cannot generally be considered. We have therefore introduced a different problem formulation for batches of tests which is tractable and can be solved using heuristics.

Finally, we have shown some preliminary results of applying the new approach to yeast species data, which has strenghten the case for this research.

Work remains to be done at a practical level, to fully implement and test all the discussed approaches. Further experimental results may allow us to establish comparisons about variations in differentiability criteria and their effect on organism differentiability. Results of applying algorithms to yeast data may be interesting to the yeast expert community. We believe the approach can be extended to other domains, and so application to other areas such as medical data may play a part in future research. Further work on this problem is in preparation and is reported in (Reynolds et al., 2003)

Notes

[1] BBSRC Grant No: 83/BIO 12037

[2] The CBS is located in Utrecht. Details can be found at http://www.cbs.knaw.nl

References

Barnett, J. A.: 1971a, 'Identifying Yeasts'. *Nature* **229**(578).

Barnett, J. A.: 1971b, 'Selection of tests for identifying yeasts'. *Nature* **232**, 221–223.

Barnett, J. A., R. W. Payne, and D. Yarrow: 2000, *Yeasts: Characteristics and identification, Third Edition.* Cambrige, UK: Cambridge University Press.

de la Iglesia, B., J. C. W. Debuse, and V. J. Rayward-Smith: 1996, 'Discovering Knowledge in Commercial Databases Using Modern Heuristic Techniques'. In: E. Simoudis, J. W. Han, and U. M. Fayyad (eds.): *Proceedings of the Second Int. Conf. on Knowledge Discovery and Data Mining.* AAAI Press.

de la Iglesia, B. and V. J. Rayward-Smith: 2001, 'The Discovery of Interesting Nuggets Using Heuristic Techniques'. In: H. A. Abbass, R. A. Sarker, and C. S. Newton (eds.): *Data Mining: a Heuristic Approach.* USA: Idea group Publishing.

Deb, K., S. Agrawal, A. Pratap, and T. Meyarivan: 2000, 'A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II'.

Debuse, J., B. de la Iglesia, C. M. Howard, and V. J. Rayward-Smith: 2000, 'Building the KDD Roadmap: A methodology for Knowledge Discovery'. In: R. Roy (ed.): *Industrial Knowledge Management.* London: Springer-Verlag, pp. 179–196.

Debuse, J. C. W. and V. J. Rayward-Smith: 1997, 'Feature Subset Selection within a Simulated Annealing Data Mining Algorithm'. *Journal of Intelligent Information Systems* **9**, 57–81.

Fonseca, C. and P. J. Fleming: 1995, 'An overview of evolutionary algorithms in multiobjective optimisation'. *Evolutionary Comp.* **3**, 1–16.

Garey, M. R. and D. S. Johnson: 1979, *Computers and intractability: a guide to the theory of NP-completeness.* New York: Freeman.

Goldberg, D. E.: 1989, *Genetic Algorithms in Search, Optimisation and Machine Learning.* Reading, Massachusetts: Addison-Wesley.

Hall, M.: 1998, 'Correlation-based Feature Selection for Machine Learning'.
Horn, J. and N. Nafpliotis: 1994, 'Multiobjective Optimisation Using the Niched Pareto Genetic Algorithm'. Technical Report Illigal Report 93005, Illinois Genetic Algorithms Laboratory, University of Illinois, Urbana, Champaign.
Hyafil, L. and R. L. Rivest: 1976, 'Constructing optimal binary decision trees is np-complete'. *Information Processing Letters* **5**, 15–17.
Jakob, W., M. Gorges-Schleuter, and B. C.: 1992, 'Applications of genetic algorithms to task planning and learning'. In: R. Manner and B. Manderick (eds.): *Parallel problem solving from Nature, 2.* North-Holland, Amsterdam: pp. 291–300.
Karp, R. M.: 1972, 'Reducibility among combinatorial problems'. In: *Complexity of Computer Communications.* New York: Plenum Press.
Kononenko, I.: 1994, 'Estimating Attributes: Analysis and Extensions of RELIEF'. In: *European Conference on Machine Learning.* pp. 171–182.
Liu, H., H. Motoda, and M. Dash: 1998, 'A Monotonic Measure for Optimal Feature Selection'. In: *European Conference on Machine Learning.* pp. 101–106.
Mann, J. W.: 1996, 'X-SAmson v1.5 Developers Manual'. *School of Information Systems Technical Report, University of East Anglia, UK.*
Osyczka, A.: 1985, 'Computer Aided Multicriterion Optimisation Method'. *Advances in Modelling and Simulation* **3**(4), 41–52.
Pankhurst, R. J. (ed.): 1975, *Systematics Association Special Volume No. 7, Biological Identification with Computers.* New York: Academic Press.
Parks, G. T. and I. Miller: 1998, 'Selective Breeding in a Multiobjective Genetic Algorithm'. In: A. E. Eiben (ed.): *Proceedings of the Fifth International Conference on Parallel Problem Solving from Nature.* Springer-Verlag.
Payne, R. W.: 1981, 'Selection Criteria for the construction of Efficient Diagnostic Keys'. *Journal of Statistical Planning and Inference* **5**, 27–36.
Payne, R. W.: 1991, 'Construction of Irredundant Test Sets'. *Applied Statistics* **40**, 213–229.
Payne, R. W.: 1992, 'The use of identification keys and diagnostic tables in statistical work'. In: *COMPSTAT 1992: Proceedings in Computational Statistics*, Vol. 2. Heidelberg, Physica-Verlag.
Payne, R. W.: 1993, 'Genkey, A program for Construction and Printing Identification Keys and Diagnostic Tables'. Technical Report m00/42529, Rothamsted Experimental Station, Harpenden, Hertfordshire.
Payne, R. W. and T. J. Dixon: 1984, 'A study of selection criteria for constructing identification keys'. In: T. Havranek, Z. Sidak, and M. Novak (eds.): *COMPSTAT 1984: Proceedings in Computational Statistics.* Vienna, Physica-Verlag.
Payne, R. W. and D. A. Preece: 1981, 'Identification keys and diagnostic tables: a review (with discussion)'. *Journal of the Royal Statistical Society* **143**, 253–292.
Payne, R. W. and C. J. Thompson: 1989, 'A study of Criteria for Constructing Identification Keys Containing Tests with Unequal Costs'. *Computational Statistics Quarterly* **1**, 43–52.
Pitt, J. I. and A. D. Hocking: 1997, *Fungi and food spoilage 2nd Edition.* London: Blackie Academic and Professional.
Quinlan, J. R.: 1993, *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann.
Quinlan, J. R.: 1996, 'Bagging, Boosting, and C4.5'. In: *Proc. of the Thirteenth National Conference on A.I.* AAAI Press/MIT Press.
Reynolds, A. P., J. L. Dicks, I. N. Roberts, J. J. Wesselink, B. de la Iglesia, V. Robert, T. Boekhout, and V. Rayward-Smith: 2003, 'Algorithms for Identifica-

tion Key Generation and Optimization with Application to Yeast Identification'. In: *Proceedings of EvoBIO-2003 LNCS, Volume 2611.* Springer. (To appear).

Schaffer, J. D.: 1985, 'Multiple objective optimisation with vector evaluated genetic algorithms'. In: J. J. Grefenstette (ed.): *Proceedings of the First International Conference on Genetic Algorithms.* San Mateo, California, pp. 93–100, Morgan Kaufmann Publishers Inc.

Srinivas, N. and K. Deb: 1994, 'Multiobjective optimisation using non-dominated sorting in genetic algorithms'. *Evolutionary Computation* **2**(3), 221–248.

Wesselink, J. J., B. de la Iglesia, S. A. James, J. L. Dicks, I. N. Roberts, and V. J. Rayward-Smity: 2002, 'Determining a unique defining DNA sequence for yeast species using hashing techniques'. *Bioinformatics* **18**(7), 1004–1010.

Willcox, W. R. and S. P. Lapage: 1972, 'Automatic Construction of Diagnostic Tables'. *Computer Journal* **15**, 263–267.

Zimmermann, H. J.: 1991, *Fuzzy Set Theory and its applications.* London: Kluwer Academic Publishers.