A genetic algorithm for the classification of natural corks

PECH-GOURG Nicolas
Group SABATÉ
Espace Tech Ulrich
66400 Céret – France
pech@sabate.fr

HAO Jin-Kao
Université d'Angers
2, bd Lavoisier
49045 Angers – France
hao@info.univ-angers.fr

Abstract

In this paper, we explore the use of genetic algorithms (GA) for a classification problem encountered in wine industry: the classification of natural corks according to the defects of their heads. In particular, we are interested in the task of optimizing the parameters of an existing cork classification program. For this purpose, we introduce a GA-based approach that searches for good combinations from a huge search space. Experiments on both artificial and real data show the high effectiveness of this approach. This effectiveness justifies the use of this approach for daily operations in a real environment.

1 INTRODUCTION

The cork is a well-known natural product in fine wine industry for its reliability and for its chemical and mechanic properties. The main advantage of a natural cork stopper is to allow a good gaseous diffusion adapted to the wine maturation. This is also the most appreciated cork by wine consumers. In cork industry, the production process of this product is composed of different steps [FOU97]. First, the cork is punched in cork planks. Then corks batches are washed and classified. The last steps consist in personifying the corks (picture, surface treatment) and to pack them up.

In this study, we are interested in the classification step. In fact, natural corks are classified according to their quality and proposed to vineyard with different prices. Like a lot of natural products, natural corks are heterogeneous. To classify them, a human expert would consider holes, cracks, colors and other features of a cork. The quality of a cork depends on the nature, the quantity, the size and the position of the defects. In the case of an automatic classification of corks, only some visual features are taken into account. In this study, we are only interested in the classification according to the visual aspect of the two heads of the cork. This operation allows separating a cork set into three categories. To obtain the necessary data for the classification, we use CCD cameras that give us pictures for each head of the cork. From these pictures we obtain numerical values. A classification program is then used to determine the class of each cork. This classification decision is taken by comparing the numerical values from the cameras against some internal parameters of the classification program. These internal parameters correspond in fact to a set of thresholds that must be determined carefully in order for the classification program to work correctly. The main difficulty is that these parameters are numerous (up to 30) and have large ranges for the possible values (up to 10.000 integer values).

The goal of this work is to explore a GA-based approach to determine these threshold values used by the classification program. We evaluate this approach on both artificially generated theoretical data and real data. We show the GA-based approach is able to find near optimal values for the classification parameters. Indeed, using these parameters values, the classification program produced excellent results for both the artificial and real data.

The paper is organized as follows. In section 2, we introduce our classification problem, followed by the presentation of a mathematical formulation of the problem in section 3. In section 4, we present our GA for determining the classification parameters. In section 5, we show detailed experimental results. Conclusions are given in the last section.

2 CORKS CLASSIFICATION AND CLASSIFICATION PARAMETERS

Four CCD cameras allow obtaining two pictures for each of the two heads of a cork. Each picture is analyzed in order to extract fifteen parameters that we will note CAM_{ij}: i represents the number of the camera (between 1 and 4) and j represents the number of the parameter (between 1 and 15). We will not explain the methods used to extract these parameters, neither the nature of the selected parameters. The problem that interests us in this study is in the following step. A classification program analyses the fifteen parameters given by each of the four cameras. The result of this program is the class of the cork. In Figure 1, we show the two heads of an example cork and the classification process working with the four corresponding numerical pictures of the two heads.

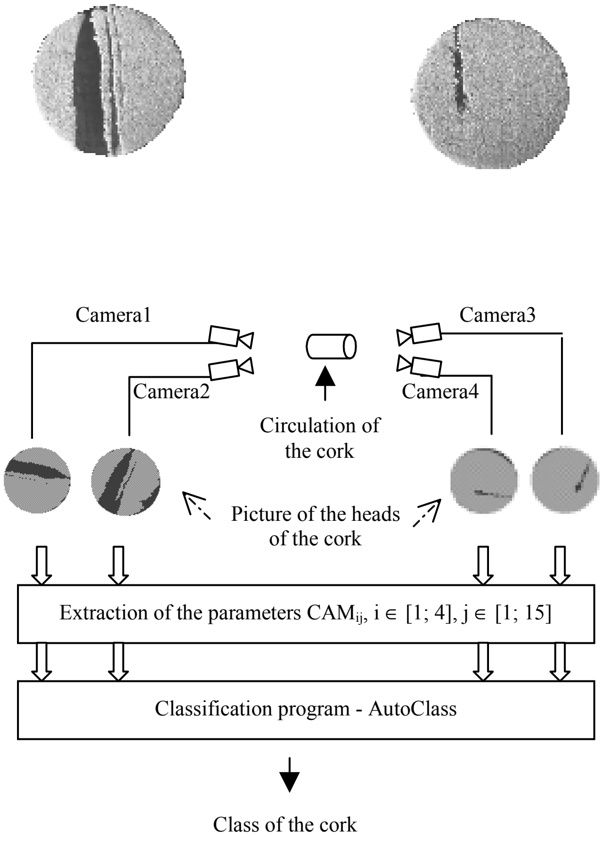
Camera1
Camera3
Camera2
Camera4
Circulation of the cork
Picture of the heads of the cork
Extraction of the parameters CAMij, i ∈ [1; 4], j ∈ [1; 15]
Classification program - AutoClass
Class of the cork

Figure 1: From the visualization to the classification of the cork

To simplify, we can say that the classification program (AutoClass) uses thirty internal parameters denoted by (P_{1i}, P_{2i}), $i \in [1; 15]$. They are the same nature as the CAM_{ij}.

The algorithm used by the classification program is quite simple: it compares the numerical values (CAM_{ij}) from the camera pictures against the classification parameters (thresholds) (P_{1i}, P_{2i}). A cork is classified to one and only one of three different classes after this comparison (Classes 1 to 3 correspond in fact to decreasing qualities).

```
IF $\forall$ (i, j) $\in$ ([1 ;4], [1; 15]), $CAM_{ij} < P_{1j}$
         THEN Class = 1;
ELSE     IF $\forall$ (i, j) $\in$ ([1 ;4], [1 ;15]), $CAM_{ij} < P_{2j}$
                  THEN Class = 2;
```

Figure 2: Classification Algorithm

Clearly, the quality of the classification parameters plays a determinant role for a good classification. A good setting of these parameters $(P_{1i}; P_{2i})$, $i \in [1; 15]$ will allow to classify a cork in the class which is the most appropriated for it according to the information given by the cameras.

3 FORMULATION

In this section, we give a formulation of our problem, which is based on the CSOP model [TSA93]. Here we identify a set of (discrete) variables V, a family of value domains for the variables, a set of constraints among some variables and a cost function to be optimized.

Variables:

$$V = \{P_{1\text{-}1}; P_{1\text{-}2}; \ldots P_{1\text{-}15}; P_{2\text{-}1}; \ldots ; P_{2\text{-}15}\}$$
$$= \{V_i ; i \in [1; 30]\}$$

The set of variables is composed of the parameters P_{1i} and P_{2i}, that are renamed as V_i, $i \in [1; 30]$.

Domains:

$$D = \{D_i \,/\, D_i = N^+, \forall\, i \in [1; 30]\}$$

Each variable V_i must take a positive and entire value. More precisely, for this study, we have $D_i = [0; 800]$ for $i \in [1;15]$, $D_{15} = [5\ 000; 15\ 000]$ and $D_{i+15} = D_i$ for $i \in [1;15]$.

Constraints:

$$C: \forall\, i \in [1; 15], V_i \leq V_{i+15}$$

This constraint is used to avoid a cork that cannot be accepted in class 2, could be accepted in class 1 (Class 1 is of higher quality). This constraint is due to the classification algorithm presented before. In fact, without this constraint, we would have: $\exists\ k \in [1; 15] \,/\, V_{k+15} < CAM_{ik} < V_k$. A cork can then be put to the class 1 (because $CAM_{ik} < V_k$), while it is rejected from class 2 (because $V_{k+15} < CAM_{ik}$). The set of the proposed constraints allows us to avoid this undesirable situation.

Cost function:

This is the sum of corks that are classified in the right way. These classified corks are those for which the class determined by the classification algorithm is the same as the known class given by the human expert. The aim is of course to maximize this function.

4 A GA-BASED RESOLUTION APPROACH

From the literature, one may find several studies concerning the automatic classification of corks by analyzing pictures of corks and by employing different classification techniques. For example, some researchers take interests in picture analysis to determine the quality of cork boards [MOL93]. Others are interested in the picture analysis and in the classification of corks with the help of artificial neuronal networks [CHA97]. In a more general context, genetic algorithms have been successfully applied to various classification-related problems [PUN93], [SIE88], [VAF91], [FAL93]. These previous studies on similar problems constitute one important factor motivating the choice of genetic algorithms for our classification problem.

Since the very beginning of the GA [HOL75], its principle becomes well known. For a comprehensive introduction, the reader is invited to consult books on the subject, for example [GOL89]. We give here only a brief remainder necessary to describe our genetic algorithm. A GA may be considered to be composed of three essential elements:

1. A set of *potential solutions* called individuals or chromosomes that will evolve during a number of iterations (generations). This set of solutions is also called *population*.
2. An *evaluation* mechanism (fitness function) that allows assessing the quality or fitness of each individual of the population.
3. An *evolution* procedure that is based on some "genetic" operators such as selection, crossover and mutation.

Crossover and Mutation

- The crossover takes two individuals to produce two new individuals. For example, the application of the well-known one-point crossover to α=abcd and ß=bbaa can produce two individuals γ=ab*aa* and η=bb*cd*.
- The mutation consists in modifying randomly a gene of an individual. A mutation of γ=ab*a*a could lead to a new individual γ=ab*e*a.

Fitness function and selection

The quality of the individuals is assessed with a fitness function. The result is a real value for each individual. The best individuals will survive and are allowed to produce new individuals.

Stop condition

The stop condition is used to determine the end of the algorithm. Well-known stop conditions are:

- a pre-defined number of generations or evaluations,
- a pre-defined value to reach for the fitness function,
- a number of generation without improvement.

Our genetic algorithm

For our problem of determining the parameters for cork classification, each individual is defined by a vector: V^i=(P^i_1, ., P^i_{10}, ., P^i_{30}), each gene corresponding to one of the thirty parameters of the problem and taking its value from its value domain (c.f. §3). A population of 40 individuals is used in this study.

The classical one-point crossover is used to generate new individuals. For the mutation, the following technique is used. Suppose we decide to mutate the kth gene V^i_k of an individual. Then the new value for the gene is determined by V^i_k + (random(1)-0.5) x V^i_k. Selection is carried out over the whole population and half of the best individuals are kept. The best individual is always record in a variable (V*) and updated each time a better solution is found. The stop condition concerns the number of generations without improvement of the best solution found so far. This number is empirically fixed at 50 generations.

To evaluate the fitness of an individual, we run the classification program AutoClass (§2) with the parameter values coded by the individual on a learning database. The learning database is composed of a set of corks with a known class number for each cork. According to the number of corks that are correctly classified, a score is assigned to the individual that is being assessed. Since we use an external program for fitness evaluation, it is clear that the evaluation constitutes the most time-consuming part of the algorithm.

In addition to these conventional mechanisms, our GA uses a diversification function: if the best individuals of the population do not evolve during 10 generations, then the whole population undergoes a mutation (each individual is mutated). This diversification function allows modifying the population more importantly than by a crossover or a classical random mutation. It helps in some cases avoid the problem of premature convergence of the population. The overall algorithm is described by the following flowchart (Figure 3).

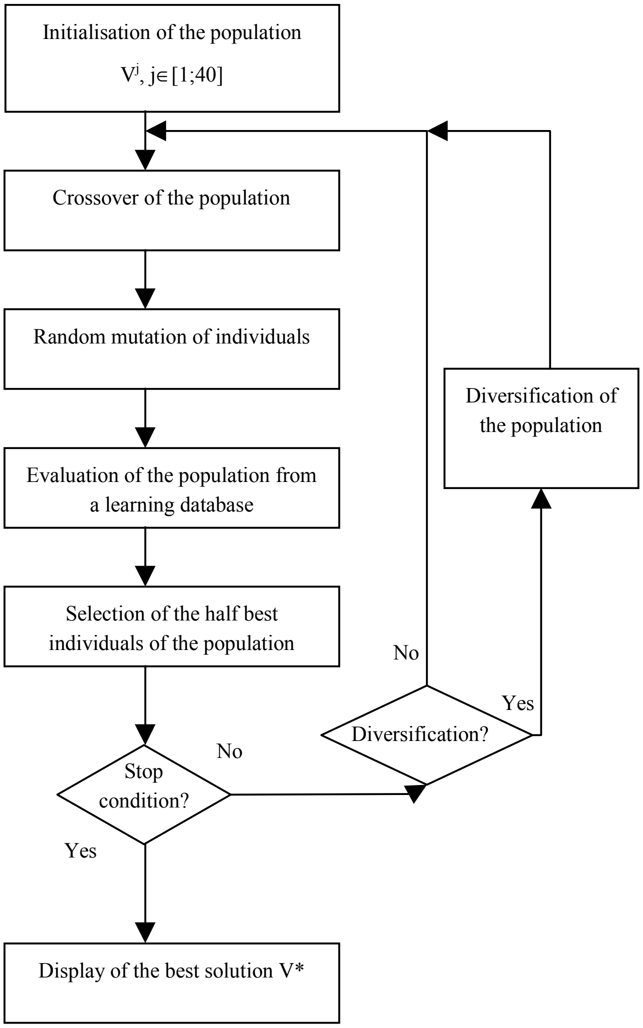
Initialisation of the population
V^j, j∈[1;40]
Crossover of the population
Random mutation of individuals
Evaluation of the population from a learning database
Selection of the half best individuals of the population
Stop condition?
No
Yes
Diversification?
No
Yes
Diversification of the population
Display of the best solution V*

Figure 3: A GA for a classification system of natural corks

5 EXPERIMENTAL RESULTS

5.1 RESULTS ON ARTIFICIAL DATA

In order to assess the approach just described, we apply the approach to a set of artificial, random data for which an optimal solution is known, that is, for each cork, we know its class. Using such a data set, we may compare directly the results of the GA with the optimal ones, consequently. These data are generated in the following way. We create a 2-dimentional N x M table with N=5000 (the number of theoretical corks) and M=61 (60 simulated numerical values that are usually given by 4 cameras plus the class of the cork).

More precisely, the first line is randomly computed and the following data are calculated from a function that takes into account the value of the cell of the first line and a random value. For each of the N lines, there are the 15 parameters given by each of the four cameras for each cork. We obtain the information on the four pictures of five thousand theoretical corks. In order to assign a cork to a class, we proceed as follows. We take randomly a combination for the thirty parameters V_i, i∈[1; 30] used by the classification program AutoClass. We run then AutoClass with these parameters to classify all the 5000 corks. In this case, we know the class of each cork and we know also the parameters necessary to find this classification (These parameters may be considered to be optimal for the classification of these corks). Now we can run our GA on these data to see whether it is able to find these optimal (or near-optimal) parameters to classify correctly all the corks of these data.

We test the program on data sets with different sizes (50, 100, 200, 500, 1000 and 5000 corks). We run 10 times the algorithm on each data set. The tests were realized on a Pentium II with 200 MHz and 64 MB of RAM. The results are given in the following table.

Table 1: Solutions found for 10 different runs on theoretical corks

N = number of corks	Case where f = N	Case where f < N	Average value of f (in %)	Average solving time for one run
50	3	7	40/50 (80%)	1 min 14 s
100	3	7	73/100 (73%)	3 min 50 s
200	2	8	162/200 (81%)	7 min 28 s
500	3	7	465/500 (93%)	26 min 10 s
1000	1	9	873/1000 (87%)	59 min 13 s
5000	2	8	4533/5000 (87%)	5 h 50 min

(population size: 40, stop condition: 50 generations without improvement)

From table 1 we observe, for example, that with 200 corks, the algorithm finds twice out of ten the optimal solution (f = N), that is, it finds twice a combination of the classification parameters V_i that allows classifying correctly all the 200 corks. On average, the algorithm leads to a right classification for 162 of 200 corks (81%). The last column indicates the average time for a run.

Let us note that the resolution time increases according to the size of the data set. This increase is due to the evaluation step that uses an external classification program (AutoClass, see §4). The more important the data set is, the higher the evaluation time is.

This experiment is very satisfactory for a practical point of view. Indeed, it shows that the algorithm is able to find the best (optimal) solution at least once out of four in the previous example. Here, we can speak of the optimal solution because it is known and we know that it is possible to reach it. With real data we will see that this is no more possible because an optimal classification is not known in advance for a given set of corks. Moreover, it is almost impossible to classify a set of corks exactly in the same way as a human expert. We discuss this issue in the next section.

5.2 A CASE STUDY ON REAL DATA: THE CLASSIFICATION OF 173 CORKS

From a visual selection realized by a human expert, 173 corks were classified according to their heads into three classes. The following table gives the result of this manual classification done by the expert.

Table 2: Classification by an expert of a batch of 173 corks

	Class 1	Class 2	Class 3	*Total*
Quantity	70	46	57	*173*
Percentage	40.5 %	26.5 %	33 %	*100 %*

We analyze the corks of each class with the four cameras to extract the sixty parameters from the cork. The data are recorded in a 61-columns table. The class determined by the human expert is indicated in the 61st column. Then, we run our algorithm to determine the 30 classification parameters V_i, $i \in [1;30]$ such that the classification is the same as that determined by the human expert.

We run twenty times the algorithm before selecting the best solution. The results are summarized in table 3.

Table 3: Results of 20 runs on 173 real corks

	Maximal value of the fitness function f (correctly classified corks for the 173 corks)	Number of generations
Run 1	130	144
Run 2	129	239
Run 3	130	158
Run 4	130	252
Run 5	130	161
Run 6	127	109
Run 7	129	135
Run 8	130	194
Run 9	129	138
Run 10	130	194
Run 11	129	174
Run 12	127	168
Run 13	130	140
Run 14	130	197
Run 15	130	212
Run 16	128	197
Run 17	130	171
Run 18	130	188
Run 19	130	204
Run 20	129	168

The next figure (figure 4) shows the typical evolution of the fitness function of the best individual with the number of generations. From the figure, we observe that the fitness of the best individuals increases quickly for the first 60 generations. Then the evolution slows down and stops around 181 with a best fitness of 130.

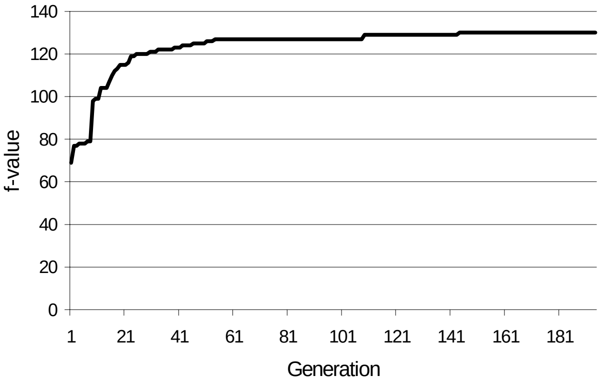

Figure 4: Evolution of the best individuals of the population

From these results, we know that the classification parameters determined by the GA allow 130 out of 173 corks to be classified as the human expert suggested. Now, we want to know exactly which cork is classified into which class. For this purpose, we take one of the best individuals (with f_{max} = 130). We re-run the classification program with the classification parameters given by the chosen individual. Applying to our 173 corks, we obtain the following results (table 4):

Table 4: Confusion matrix for a total of 173 corks.

Expert \ Machine	Class 1	Class 2	Class 3	Total
Class 1	**61**	8	1	*70*
Class 2	10	**23**	13	*46*
Class 3	7	4	**46**	*57*

Classified in the right class: 61 + 23 + 46 = 130

Satisfaction Percentage: 75.1%

From table 4, we can see that on the 70 corks that are classified by the expert in the class 1, 61 of them are classified by the classification system in class 1, 8 in class 2, and 1 in class 3. For the 173 corks, the algorithm leads to a classification that has an overlap of 75.1% with that of the human expert.

If we compare these results with those obtained on theoretical corks (§5.1), we may conclude that the results on real data are "less good". Two factors can explain the difference between these two experiments. The first one is due to the classification made by the human expert (cf. table 2). Just like we realized a confusion matrix between a human expert and a classification program (cf. table 4), we also could realize a confusion matrix between two experts or with the same expert but in different conditions. Without any doubts, the traces of the matrix would never be equal to the number of corks to be classified. This result is well known in cork industry and certainly also in other domains that use the human intervention of man to classify products.

The second factor is a more bothering one that is related to the classification algorithm currently used (AutoClass). The data themselves we use may not allow classifying correctly the set of corks. Take an example with two variables, noted Var_1 and Var_2, and two classes to be separated: the circles and the triangles (cf. figure 5). There is an obvious manner to separate these elements: the straight θ. However, the classification algorithm AutoClass is not able to separate these elements by using ψ and φ (perpendicular to the axes represented by the variables). In the case presented here, there is no way to separate the two classes with ψ and φ.

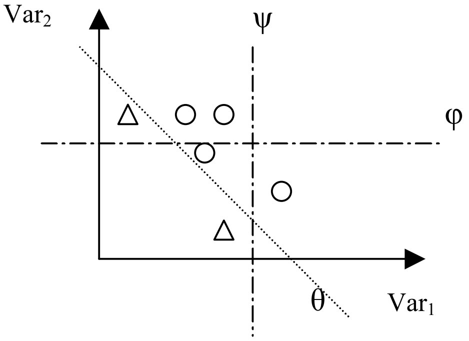

Figure 5: Separation of classes

These two factors explain the difference between the quality of theoretical data and the tested real data.

Let us mention that other tests have been carried out on very large set of non-classified corks (up to 15 000 corks). Assessed by human expert, the classification results on these real data are considered to the best one known today for the daily industrial classification task. For this reason, the system is currently used in daily operation.

6 CONCLUSIONS AND FUTURE WORK

The classification of natural corks is a very important topic in wine industry. In this paper, we have studied a parameter optimization problem for an automatic classification system. The problem involves thirty variables with a huge number (up to 10 000) of possible values for these parameters. To solve the problem, we have developed a GA-based approach to search for good combinations for the thirty parameters of the problem. The proposed approach has been evaluated on both (supervised) artificial data and real data. These evaluations have led to highly satisfactory and concluding results on the tested data. Moreover, results on unsupervised data were favorably approved by human expert and were the best ones known.

The analysis of results showed that it would still be possible to improve the effectiveness of the classification system by modifying other steps of the classification process (including the classification program used currently). One possibility would be to use a GA to find more pertinent classification rules. We studied in this paper the classification only according to the defects of the heads of the cork. Classification is also done using defects of boards of the cork. We would use the approach proposed in the paper to this kind of classification. Finally, we plan to apply the proposed approach to other classification problems encountered in wine industry. For example, for champagnes corks, one distinguishes even more classification steps: the classification of the two slices before pasting them, and the classification of corks according to the specification of customers (who become our expert!).

Acknowledgements

This work was realized within the framework of a CIFRE convention between the group SABATÉ SA and the University of Angers. The authors would like to thank the "Agence Nationale de la Recherche Technique" (ANRT) for the financial support. Special thanks go to M. Bussac from SABATÉ S.A. for his help for this research and M. Hyppolyte and M. Nougalliat from VGA S.A. for their technical support about cork classification. We thanks also the reviewers of the paper for their comments.

References

[CHA97] CHANG J., HAN G., VALVERDE J.M., GRISWOLD N.C., DUQUE-CARRILLO J.F., SANCHEZ-SINENCIO E., Cork quality classification system using a unified image processing and fuzzy-neural network methodology*, IEEE Transactions on neural networks*. Vol. 8 No. 4, pages 964-974, 1997.

[FAL93] FALKENAUER E., GASPART P., Creating part families with a grouping genetic algorithm, *International Symposium on Intelligent Robotics*, India, 1993.

[FOU97] FOUCAULT V., Code international des pratiques bouchonnières, *Confédération européenne du liège*, 1997.

[GOL89] GOLDBERG D.E., *Genetic algorithm in search, optimization and machine learning*, Addison-Wesley Publishing Campany, Inc, 1989.

[HOL75] HOLLAND J.H., *Adaptation in natural and artificial systems*, The University of Michigan Press, 1975.

[MOL93] MOLINAS M., CAMPOS M., Aplicacion del analisis digital de imagenes al estudio de la calida del corcho, *Congreso forestal espanol*, Lourizan, Ponencias y Comunicaciones, Vol. 6, pages 347-352, 1993.

[PUN93] PUNCH W.F., GOODMAN E.D., PEI M., CHIA-SHUN L., HOVLAND P., ENBODY R., Further research on feature selection and classification using genetic algorithms, *ICGA93*, pages 557-564, 1993.

[SIE88] SIEDLECKI W., SKLANSKY J., On automatic feature selection, *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 2, No. 2, pages 197-220, 1988.

[TSA93] TSANG E., *Foundations of constraint satisfaction*, Academic Press, 1993.

[VAF92] VAFAIE H., JONG DE K., Genetic algorithms as a tool for feature selection in machine learning, *Proceedings of the Intl. Conf. on Tools with AI*, Arlington, VA, pages 200-204. IEEE CS Press, 1992 .