Feature Selection Using Fuzzy Objective Functions

Susana M. Vieira[1] João M. C. Sousa[1] Uzay Kaymak[2]

1. Instituto Superior Técnico, Technical University of Lisbon
Lisbon, Portugal
2.Econometric Institute, Erasmus School of Economics, Erasmus University of Rotterdam
Rotterdam, The Netherlands
Email: susana@dem.ist.utl.pt, jmsousa@ist.utl.pt, u.kaymak@ieee.org

Abstract *— One of the most important stages in data preprocessing for data mining is feature selection. Real-world data analysis, data mining, classification and modeling problems usually involve a large number of candidate inputs or features. Less relevant or highly correlated features decrease, in general, the classification accuracy, and enlarge the complexity of the classifier. Feature selection is a multi-criteria optimization problem, with contradictory objectives, which are difficult to properly describe by conventional cost functions. The use of fuzzy decision making may improve the performance of this type of systems, since it allows an easier and transparent description of the different criteria used in the feature selection process. In previous work an ant colony optimization algorithm for feature selection was presented, which minimizes two objectives: number of features and classification error. Two pheromone matrices and two different heuristics are used for each objective. In this paper, a fuzzy objective function is proposed to cope with the difficulty of weighting the different criteria involved in the optimization algorithm.*

Keywords— Feature selection, fuzzy decision functions, ant colony optimization.

1 Introduction

Feature selection has been an active research area in data mining, pattern recognition and statistics communities. The main idea of feature selection is to choose a subset of available features, by eliminating features with little or no predictive information and also redundant features that are strongly correlated. Many practical pattern classification tasks (e.g., medical diagnosis) require learning of an appropriate classification function that assigns a given input pattern (typically represented by using a vector of feature values) to one of a set of classes. The choice of features used for classification has an impact on the accuracy of the classifier and on the time required for classification.

The challenge is selecting the minimum subset of features with little or no loss of classification accuracy. The feature subset selection problem consists of identifying and selecting a useful subset of features from a larger set of often mutually redundant, possibly irrelevant, features with different associated importance [1].

Like many design problems, feature selection problem, is characterized by multiple objectives, where a trade-off amongst various objectives must be made, leading to under or over-achievement of different objectives. Moreover, some flexibility may be present for specifying the constraints of the problem. Furthermore, some of the objectives in decision making may be known only approximately. Fuzzy set theory provides ways of representing and dealing with flexible or soft constraints, in which the flexibility in the constraints can be exploited to obtain additional trade-off between improving the objectives and satisfying the constraints.

Various fuzzy optimization methods have been proposed in the literature in order to deal with different aspects of soft constraints. In one formulation of fuzzy optimization due to Zimmermann [2], which is used in the rest of this paper, concepts from Bellman and Zadeh model of fuzzy decision making [3] are used for formulating the fuzzy optimization problem.

In this paper is used a feature selection algorithm which is based in ant colony optimization. The ant feature selection algorithm uses two cooperative ant colonies, which are used to cope with two different objectives. The two objectives we consider are minimizing the number of features and minimizing the classification error. Two pheromone matrices and two different heuristics are used for each objective. These goals are translated into fuzzy sets.

The paper is organized as follows. Section 2 presents a brief description of fuzzy optimization. A brief consideration of the fuzzy models we use is presented in Section 3. The ACO feature selection algorithm is presented in Section 4. In Section 5 the results are presented and discussed. Some conclusions are drawn in Section 6 and the possible future work is discussed.

2 Fuzzy Optimization

Fuzzy optimization is the name given to the collection of techniques that formulate optimization problems with flexible, approximate or uncertain constraints and goals by using fuzzy sets. In general, fuzzy sets are used in two different ways in fuzzy optimization.

1. To represent uncertainty in the constraints and the goals (objective functions).
2. To represent flexibility in the constraints and the goals.

In the first case, fuzzy sets represent generalized formulations of intervals that are manipulated according to rules similar to interval calculus by using the α-cuts of fuzzy sets. In the second case, fuzzy sets represent the degree of satisfaction of the constraints or of the aspiration levels of the goals, given the flexibility in the formulation. Hence, the constraints (and the goals) that are essentially crisp are assumed to have some flexibility that can be exploited for improving the optimization objective. This framework is suitable for the representation of interaction and possible trade-off amongst the constraints and the objectives of the optimization [4].

ISBN: 978-989-95079-6-8

2.1 General formulation

The general formulation for *fuzzy optimization* in the presence of flexible goals and constraints is given by

$$\begin{array}{ll} \underset{\mathbf{x}\in X}{\text{fuzzy maximize}} & [f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x})] \\ \text{subject to} & g_i(\mathbf{x}) \tilde{\leq} 0, \quad i = 1, 2, \ldots, m. \end{array} \tag{1}$$

In (1), the tilde sign denotes a fuzzy satisfaction of the constraints. The sign $\tilde{\leq}$ thus denotes that $g_i(\mathbf{x}) \leq 0$ can be satisfied to a degree smaller than 1. The fuzzy maximization corresponds to achieving the highest possible aspiration level for the goals $f_1(\mathbf{x})$ to $f_n(\mathbf{x})$, given the fuzzy constraints to the problem. This optimization problem can be solved by using the approach of Bellman and Zadeh to fuzzy decision making [3].

Consider a decision making problem where the decision alternatives are $\mathbf{x} \in X$. A fuzzy goal F_j, $j = 1, 2, \ldots, n$ is a fuzzy subset of X. Its membership function $F_j(\mathbf{x})$, with $F_j : X \rightarrow [0, 1]$ indicates the degree of satisfaction of the decision goal by the decision alternative $\mathbf{x}$. Similarly, a number of fuzzy constraints G_i, $i = 1, 2, \ldots, m$ can be defined as fuzzy subsets of X. Their membership functions $G_i(\mathbf{x})$, denote the degree of satisfaction of the fuzzy constraint G_i by the decision alternative $\mathbf{x} \in X$. According to Bellman and Zadeh's fuzzy decision making model, the fuzzy decision D is defined as the confluence of the fuzzy goals and constraints, i.e.

$$D(\mathbf{x}) = F_1(\mathbf{x}) \circ \cdots \circ F_n(\mathbf{x}) \circ G_1(\mathbf{x}) \circ \cdots \circ G_m(\mathbf{x}), \tag{2}$$

where $\circ$ denotes an aggregation operator for fuzzy sets. Since the goals and the constraints must be satisfied simultaneously, Bellman and Zadeh proposed to use an intersection operator, i.e. a fuzzy t-norm for the aggregation. The optimal decision alternative $\mathbf{x}^*$ is then the argument that maximizes the fuzzy decision, i.e.

$$\mathbf{x}^* = \arg\max_{\mathbf{x}\in X} D(\mathbf{x}). \tag{3}$$

The optimization problem is then defined by

$$\max_{\mathbf{x}\in X} [F_1(\mathbf{x}) \wedge \cdots \wedge F_n(\mathbf{x}) \wedge G_1(\mathbf{x}) \wedge \cdots \wedge G_m(\mathbf{x})]. \tag{4}$$

Note that both the goals and the constraints are aggregated. Hence, the goals and the constraints are treated equivalently, which is why the model is said to be symmetric. The symmetric model is not always appropriate, however, since the aggregation of the goals and the constraints may have different requirements. Often, for example, some trade off amongst the goals is allowed or may even be desirable, which may be modeled by an averaging operation. The constraints, however, should not be violated, i.e. their aggregation must be conjunctive. In that case, the goals and the constraints cannot be combined uniformly by using a single aggregation operator. In the simplest case, the goals must be combined by using one operator and the constraints must be combined by using another operator. The aggregated results must then be combined at a higher level by using a third aggregation operator, which has to be conjunctive (i.e. both the aggregated goals and the aggregated constraints should be satisfied).

Clearly, the above formulation of fuzzy optimization is closely related to the penalty function methods known from classical optimization theory. The aggregated goals correspond to an overall objective function, which is maximized. The constraints are added to this objective function by using fuzzy t-norms, which is similar to the addition of a penalty function to an optimization objective function in classical optimization. After combining the objectives and the constraints, the resulting optimization is unconstrained, but possibly non-convex. Furthermore, gradient descent methods may not be suitable for the maximization due to possible and likely discontinuity in the first derivative of the final aggregated function. Derivative-free search and optimization algorithms such as simulated annealing, evolutionary algorithms or other bio-inspired algorithms, such as ant colony optimization, can be used to solve this type of optimization problems.

2.2 Weighted aggregation in fuzzy optimization

Weighted aggregation has been used quite extensively especially in fuzzy decision making, where the weights are used to represent the relative importance that the decision maker attaches to different decision criteria (goals or constraints). Almost always an averaging operator has been used for the weighted aggregation, such as the generalized means [5].

The averaging operators are suitable for modeling compensatory aggregation. They are not suitable, however, for modeling simultaneous satisfaction of aggregated criteria. Since the goal in fuzzy optimization is the simultaneous satisfaction of the optimization objectives and the constraints, t-norms must be used to model the conjunctive aggregation. In order to use the weighted aggregation in fuzzy optimization, weighted aggregation using t-norms must thus be considered.

The axiomatic definition of t-norms does not allow for weighted aggregation. In order to obtain a weighted extension of t-norms, some of the axiomatic requirements must be dropped. Especially the commutativity and the associativity properties must be dropped, since weighted operators are by definition not commutative.

Weighted t-norms. In [4], the authors used weighted counterparts of several t-norms for fuzzy optimization. The specific operators considered are the weighted extension of the product t-norm given by

$$D(\mathbf{x}, \mathbf{w}) = \prod_{i=1}^{m} [G_i(\mathbf{x})]^{w_i}, \tag{5}$$

the extension of the Yager t-norm given by

$$D(\mathbf{x}, \mathbf{w}) = \max\left(0, 1 - \sqrt[s]{\sum_{i=1}^{m} w_i (1 - G_i(\mathbf{x}))^s}\right), \quad s > 0. \tag{6}$$

The term fuzzy optimization in the remainder of this paper also refers to a formulation in terms of the flexibility of the goals.

3 Fuzzy Models for Classification

Fuzzy modeling for classification, is a technique that allows an approximation of nonlinear systems when there is none or few knowledge of the system to be modeled [6]. The fuzzy modeling approach has several advantages when compared to

ISBN: 978-989-95079-6-8

other nonlinear modeling techniques. In general, fuzzy models can provide a more transparent model and can also give a linguistic interpretation in the form of rules, which is appealing when dealing with classification systems. Fuzzy models use rules and logical connectives to establish relations between the features defined to derive the model. This paper uses Takagi-Sugeno (TS) fuzzy models [7], which consist of fuzzy rules where each rule describes a local input-output relation, typically in an affine form. The affine form of a TS model is given by:

$$R_i : \textbf{If } x_1 \text{ is } A_{i1} \textbf{ and } \ldots \textbf{ and } x_n \text{ is } A_{in} \textbf{ then } y_i = a_{i1}x_1 + \ldots + a_{in}x_n + b_i\,, \quad (7)$$

where $i = 1, \ldots, K$, K denotes the number of rules in the rule base, R_i is the i^{th} rule, $\mathbf{x} = [x_1, \ldots, x_n]^T$ is the antecedent vector, n is the number of features, $A_{i1}, \ldots, A_{in}$ are fuzzy sets defined in the antecedent space, y_i is the output variable for rule i, $\mathbf{a}_i$ is a parameter vector and b_i is a scalar offset. The consequents of the affine TS model are hyperplanes in the product space of the inputs and the output. The model output, y, can then be computed by aggregating the individual rules contribution:

$$y = \frac{\sum_{i=1}^{K} \beta_i y_i}{\sum_{i=1}^{K} \beta_i}, \quad (8)$$

where β_i is the degree of activation of the ith rule:

$$\beta_i = \prod_{j=1}^{n} \mu_{A_{ij}}(x_j), \quad (9)$$

and $\mu_{A_{ij}}(x_j) : \mathbb{R} \rightarrow [0, 1]$ is the membership function of the fuzzy set A_{ij} in the antecedent of R_i.

The performance criterion used to evaluate the fuzzy model is the classification accuracy C_a, given by the percentage of correct classifications:

$$C_a = \frac{(N_n - N_e)}{N_n} \times 100\%, \quad (10)$$

where N_n is the number of used samples and N_e is the number of classification errors in test samples (misclassifications).

4 Ant Feature Selection

Ant algorithms were first proposed by Dorigo [8] as a multi-agent approach to difficult combinatorial optimization problems, such as traveling salesman problem, quadratic assignment problem or supply chain management [9, 10]. The ACO methodology is an optimization method suited to find minimum cost paths in optimization problems described by graphs [11]. This paper presents a new implementation of ACO applied to feature selection, where the best number of features is determined automatically. In this approach, two objectives are considered: minimizing the number of features and minimizing the classification error. Two cooperative ant colonies optimize each objective. The first colony determines the number (cardinality) of features and the second selects the features based on the cardinality given by the first colony. Thus, two pheromone matrices and two different heuristics are used. A novel approach for computing a heuristic value is proposed to determine the cardinality of features. The heuristic value is computed using the Fisher discriminant criterion for feature selection [12], which ranks the features giving them a given relative importance and it is described in more detail in section 4.2.3. The best number of features is called *features cardinality* N_f. The determination of the *features cardinality* is addressed in the first colony sharing the same minimization cost function with the second colony, which in this case aggregates both the maximization of the classification accuracy and the minimization of the features cardinality. Hence, the first colony determines the size of the subsets of the ants in the second colony, and the second colony selects the features that will be part of the subsets.

The algorithm used in this paper deals with the feature selection problem as a multi–criteria problem with a single objective function. Therefore, a pheromone matrix is computed for each criterion, and different heuristics are used.

The objective function of this optimization algorithm aggregate both criteria, the minimization of the classification error rate and the minimization of the features cardinality:

$$J^k = w_1 \frac{N_e^k}{N_n} + w_2 \frac{N_f^k}{n} \quad (11)$$

where $k = 1, \ldots, g$, N_n is the number of used data samples and n is the total number of features. The weights w_1 and w_2 are selected based on experiments.

To evaluate the classification error, a fuzzy classifier is built for each solution following the procedure described in Section 3.

4.1 Fuzzy decision function

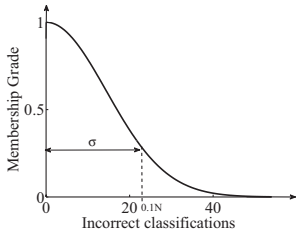
The objective function (11) can be interpreted as follows. The term containing the predicted errors indicates that these should be minimized, while the term containing the size of the features subset used indicates that the number of features should be reduced. Hence, minimizing the output errors and the size of the features subset can be regarded as forcing the model to be less complex and maintain a good performance at the same time. The parameters containing the weights, w_1 and w_2, can be changed so that the objective function is modified in order to lead to a solution where the accuracy of the model is more important or to a much simpler model where the reduction of the number of the features is imperative. This balance always depends on the application.

When fuzzy multicriteria decision making is applied to determine the objective function, additional flexibility is introduced. Each criterion ζ_i is described by a fuzzy set, where $i = 1, \ldots, T$, stands for the different criteria defined. Fuzzy criteria can be described in different ways. The most straightforward and easy way is just to adapt the criteria defined for the classical objective functions. Fig. 1 shows examples of general membership functions that can be used for the error N_e^k and for the features cardinality N_f^k, with $k = 1, \ldots, g$. In this example, the minimization of the classification error is represented by an exponential membership function, given by

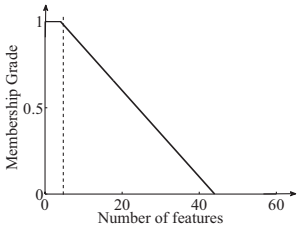
$$\mu_e^k = \exp^{-\left(\frac{N_e^k - a}{2\sigma}\right)^2} \quad (12)$$

This well-known function has the nice property of never reaching the value zero, and the membership value is still quite considerable for an error of 10%. Therefore, this criterion is

ISBN: 978-989-95079-6-8



(a) Membership function for the number of incorrect classifications.



(b) Membership function for the number of selected features

Figure 1: Membership functions for the feature selection goals.

considered to be a fuzzy goal. The features cardinality N_f can be represented, for example, by a trapezoidal membership function around zero, as shown in Fig. 1b. A reduced number of features is considered to be a desired outcome of the optimization algorithm. The membership function is defined so that for a low number of features the membership degree is one and linearly decreases to zero. The membership degree should be zero outside the maximum number of features available. The parameters defining the range of the trapezoidal membership function are application dependent. Sometimes it is convenient to make the upper limit of the membership function much lower than the maximum number of features allowed, specially if a very large number of features is being tested. In general, all the parameters of the different membership functions are application dependent. However, it is possible to derive some tuning guidelines, as will be described here. The membership functions quantify how much the system satisfies the criteria given a particular feature subset solution, bringing various quantities into a unified domain. The use of the membership functions introduces additional flexibility to the goals, and it leads to increased transparency as it becomes possible to specify explicitly what kind of solution is preferred. For instance, it becomes easier to penalize more severely a subset of features that have larger classification errors. Or if we prefer a solution with less features, a higher number of features can also be penalized.

Note that there is no need to scale the several parameters and as in (11), when fuzzy objective functions are used, because the use of membership functions introduce directly the normalization required. For this particular aspect, this feature reduces the effort on defining the objective function, when compared to classical objective functions. After the membership functions have been defined, they are combined by using a decision function, such as a parametric aggregation operator from the fuzzy sets theory (see Section 2).

Table 1: List of symbols.

Variable	Description
General	
n	Number of features
N	Number of samples
N_n	Number of samples used for validation
I	Number of iterations
K	Number of rules/clusters of the fuzzy model
N_c	Number of existing classes in database
g	Number of ants
$\mathbf{x}$	Set with all the features
$\mathbf{w}$	Subset of features selected to build classifiers
J^k	Cost of the solution for each ant k
J^q	Cost of the winner ant q
Ant colony for cardinality of features	
N_f	Features cardinality (number of selected features)
$N_f(k)$	Features cardinality of ant k
I_n	Number of iterations with same feature cardinality
α_n	Pheromone weight of features cardinality
β_n	Heuristic weight of features cardinality
τ_n	Pheromone trails for features cardinality
η_n	Heuristic of features cardinality
ρ_n	Evaporation of features cardinality
Γ_n^k	Feasible neighborhood of ant k (features cardinality availability)
$\mathcal{Q}_i$	Amount of pheromone laid in the features cardinality of the best solution
Ant colony for selecting subset of features	
$L_f^k(t)$	Feature subset for ant k at tour t
α_f	Pheromone weight of features
β_f	Heuristic weight of features
τ_f	Pheromone trails for feature selection
η_f	Heuristic of features
ρ_f	Evaporation of features
Γ_f^k	Feasible neighborhood of ant k (features availability)
$\mathcal{Q}_j$	Amount of pheromone laid in the features of the best solution

4.2 Algorithm description

Let $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ be the set of given n features, and $\mathbf{w} = [w_1, w_2, \ldots, w_{N_f}]^T$, be a subset of features where $(\mathbf{w} \subseteq \mathbf{x})$. It is desirable that $N_f << n$. Table 1 describes the variables used in the algorithm.

4.2.1 Probabilistic rule

Consider a problem with N_f nodes and two colonies of g ants. First, g ants of the first colony randomly select the number of nodes N_f to be used by the g ants of the second colony. The probability that an ant k chooses the features cardinality $N_f(k)$ is given by

$$p_i^k(t) = \frac{[\tau_{n_i}]^{\alpha_n} \cdot [\eta_{n_i}]^{\beta_n}}{\sum_{l \in \Gamma_n^k} [\tau_{n_l}]^{\alpha_n} \cdot [\eta_{n_l}]^{\beta_n}} \tag{13}$$

where τ_{n_i} is the pheromone concentration matrix and η_{n_i} is the heuristic function matrix, for path (i). The values of the

ISBN: 978-989-95079-6-8

pheromone matrix are limited to $[\tau_{n_{min}}, \tau_{n_{max}}]$, with $\tau_{n_{min}} = 0$ and $\tau_{n_{max}} = 1$. Γ_n^k is the feasible neighborhood of ant k (available number of features to be selected), which acts as the memory of the ants, and contains all the trails that the ants have not passed and can be chosen. The parameters α_n and β_n measure the relative importance of trail pheromone and heuristic knowledge, respectively.

After all the g ants from the first colony have chosen the features cardinality $N_f(k)$, each ant k from the second colony select $N_f(k)$ features (nodes). The probability that an ant k chooses feature j as the next feature to visit is given by

$$p_j^k(t) = \frac{[\tau_{f_j}(t)]^{\alpha_f} \cdot [\eta_{f_j}]^{\beta_f}}{\sum_{l \in \Gamma_f^k} [\tau_{f_l}(t)]^{\alpha_f} \cdot [\eta_{f_l}]^{\beta_f}} \quad (14)$$

where τ_{f_j} is the pheromone concentration matrix and η_{f_j} is the heuristic function matrix for the path (j). Again, the pheromone matrix values are limited to $[\tau_{f_{min}}, \tau_{f_{max}}]$, with $\tau_{f_{min}} = 0$ and $\tau_{f_{max}} = 1$. Γ_f is the feasible neighborhood of ant k (available features), which contains all the features that the ants have not selected and can be chosen. Again, the parameters α_f and β_f measure the relative importance of trail pheromone and heuristic knowledge, respectively.

4.2.2 Updating rule

After a complete tour, when all the g ants have visited all the $N_f(k)$ nodes, both pheromone concentration in the trails are updated by

$$\tau_{n_i}(t+1) = \tau_{n_i}(t) \times (1 - \rho_n) + \Delta\tau_{n_i}(t) \quad (15)$$

$$\tau_{f_j}(t+1) = \tau_{f_j}(t) \times (1 - \rho_f) + \Delta\tau_{f_j}(t) \quad (16)$$

where $\rho_n \in [0, 1]$ is the pheromone evaporation of the features cardinality , $\rho_f \in [0, 1]$ is the pheromone evaporation of the features and $\Delta\tau_{n_i}$ and $\Delta\tau_{f_j}$ are the pheromone deposited on the trails (i) and (j), respectively, by the ant q that found the best solution J^q for this tour:

$$\Delta\tau_{n_i}^q = \begin{cases} \mathcal{Q}_i & \text{if node } (i) \text{ is used by the ant } q \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

$$\Delta\tau_{f_j}^q = \begin{cases} \mathcal{Q}_j & \text{if node } (j) \text{ is used by the ant } q \\ 0 & \text{otherwise} \end{cases} \quad (18)$$

The number of nodes $N_f(k)$ that each ant k has to visit on each tour t is only updated every I_n tours (iterations), in order to allow the search for the best features for each features cardinality N_f. The algorithm runs I times. Both colonies share the same cost function given in (11).

4.2.3 Heuristics

The heuristic value used for each feature (ants visibility) for the second colony, is computed as

$$\eta_{f_j} = 1/N_{e_j} \quad (19)$$

for $j = 1, \ldots, n$. For the features cardinality (first colony), the heuristic value is computed using the

Fisher discriminant criterion for feature selection [12]. Considering a classification problem with two possible classes, class 1 and class 2, the Fisher discriminant criterion is described as

$$F(i) = \frac{|\mu_1(i) - \mu_2(i)|^2}{\sigma_1^2 + \sigma_2^2} \quad (20)$$

Algorithm 1 Ant Feature Selection

```
/*Initialization*/
set the parameters ρ_f, ρ_n, α_f, α_n, β_f, β_n, I, I_n, g.
for t = 1 to I do
  for k = 1 to g do
    Choose the subset size N_f(k) of each ant k using (13)
  end for
  for l = 1 to I_n do
    for k = 1 to g do
      Build feature set L_f^k(t) by choosing N_f(k) features using (14)
      Compute the fuzzy model using the L_f^k(t) path selected by ant k
      Compute the cost function J^k(t)
      Update J^q
    end for
    Update pheromone trails τ_{n_i}(t+1) and τ_{f_j}(t+1), as defined in (15) and (16).
  end for
end for
```

where $\mu_1(i)$ and $\mu_2(i)$ are the mean values of feature i for the samples in class 1 and class 2, and σ_1^2 and σ_2^2 are the variances of feature i for the samples in class 1 and 2. The score aims to maximize the between-class difference and minimize the within-class spread. Other currently proposed rank-based criteria generally come from similar considerations and show similar performance [12]. Since our goal is to work with several classification problems, which can contain two or more possible classes, a one versus-all strategy is used to rank features. Thus, for a C-class prediction problem, a particular class is compared with the other $C-1$ classes that are considered together. The features are weighted according to the total score summed over all C comparisons:

$$\sum_{j=1}^{C} F_j(i), \quad (21)$$

where $F_j(i)$ denotes the Fisher discriminant score for the i^{th} feature at the j^{th} comparison. Algorithm 1 presents the description of the ant feature selection algorithm.

5 Experimental Results

The effectiveness of the proposed method is applied to a data set taken from the UCI repository [13].

The classification error rates of the used classifiers are obtained by performing 10 independent runs. The data set is divided in test and training sets. The experimental results are presented as the best, the worst and the mean value of correct classifications C_a out of ten independent runs. These ten runs were obtained using the same test data set.

Wine data set The wine data set is a widely used classification data available online in the repository of the University of California [13], and contains the chemical analysis of 178 wines grown in the same region in Italy, derived from three different cultivars. Thirteen continuous attributes are available for classification: alcohol, malic acid, ash, alkalinity of ash, magnesium, total phenols, flavanoids, nonflavanoids phenols, proanthocyanism, color intensity, hue, OD280/OD315 of

ISBN: 978-989-95079-6-8

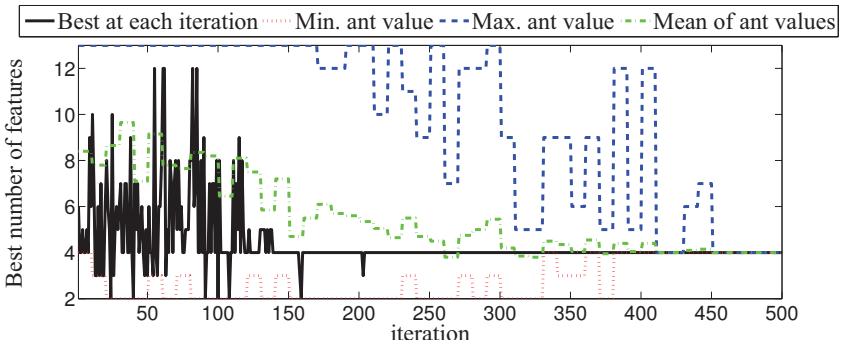


Figure 2: Convergence of the best number of feature during an algorithm run.

Table 2: Classification rates on the Wine data.

Method	**Number of features**	**Classification accuracy** (%)		
		Max.	**Mean**	**Min.**
AFS classic obj. func.	4-8	100	99.8	98.8
AFS fuzzy obj. func.	4	100	99.7	98.6

dilluted wines and proline. This data set has 13 features, three different classes and 178 samples. The AFS algorithm with fuzzy objective function is applied to select the relevant features within the wine classification data set and is compared to the same algorithm with a non fuzzy objective function.

As can be seen in Table 2, the obtained results are better than those in feature selection with classical objective function, once fewer features are selected and because the algorithm always converges to the same number of features.

An example of the process of the ant feature selection with fuzzy objectives searching for optimal solutions for wine data set is given in Fig. 2, where it is possible to observe how all the ants converge to the same solution.

6 Conclusions

A fuzzy objective function for ant feature selection is proposed in this paper. The problem is divided into two objectives: choosing the features cardinality and selecting the most relevant features. The feature selection algorithm uses fuzzy classifiers. The proposed algorithm was applied to a well known classification database that is considered a benchmark. The performance of the proposed algorithm was compared to previous works. The ant based feature selection algorithm yielded similar or better classification rates and the convergence of the solution is better than the previous approach.

In the near future, the proposed feature selection algorithm will be applied to classification problems with a larger number of features.

Acknowledgements

This work is supported by the Portuguese Government and FEDER under the programs: Programa de financiamento Plurianual das Unidades de I&D da FCT (POCTI-SFA-10-46-IDMEC) and by the FCT grant SFRH/25381/2005, Fundação para a Ciência e a Tecnologia, Ministério do Ensino Superior, da Ciência e da Tecnologia, Portugal.

References

[1] H. Motoda H. Liu. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, 1998.

[2] H. J. Zimmermann. Description and optimization of fuzzy systems. *International Journal of General Systems*, 2:209–215, 1976.

[3] R. E. Bellman and L. A. Zadeh. Decision-making in a fuzzy environment. *Management Science*, 17(4):141–164, 1970.

[4] U. Kaymak and J.M. Sousa. Weighted constraint aggregation in fuzzy optimization. *Constraints*, 8(1):61–78, January 2003.

[5] U. Kaymak and H. R. van Nauta Lemke. A sensitivity analysis approach to introducing weight factors into decision functions in fuzzy multicriteria decision making. *Fuzzy Sets and Systems*, 97(2):169–182, July 1998.

[6] L. F. Mendonça, S. M. Vieira, and J. M. C. Sousa. Decision tree search methods in fuzzy modeling and classification. *International Journal of Approximate Reasoning*, 44(2):106–123, 2007.

[7] T. Takagi and M. Sugeno. Fuzzy identification of systems and its applications to modelling and control. *IEEE Transactions on Systems, Man and Cybernetics*, 15(1):116–132, 1985.

[8] M. Dorigo. *Optimization, Learning and Natural Algorithms (in Italian)*. PhD thesis, 1992.

[9] C. A. Silva, J. M. C. Sousa, and T. A. Runkler. Rescheduling and optimization of logistic processes using GA and ACO. *Engineering Applications of Artificial Intelligence*, 21(3):343–352, 2007.

[10] C. A. Silva, J. M. C. Sousa, T. A. Runkler, and J. M. G. Sá da Costa. Distributed optimization of a logistic system and its suppliers using ant colony optimization. *International Journal of Systems Science*, 37(8):503–512, 2006.

[11] Marco Dorigo, M. Birattari, and T. Stützle. Ant colony optimization. *IEEE Computational Intelligence Magazine*, 1(4):28–39, 2006.

[12] R. O. Duda, P. E. Hart, and D. G Stork. *Pattern Classification, Second Edition*. Wiley–Interscience Publication, 2001.

[13] A. Asuncion and D.J. Newman. UCI machine learning repository, 2007.

ISBN: 978-989-95079-6-8