Abstract—This paper proposes a novel hybrid algorithm for feature selection based on a binary ant colony and SVM. The final subset selection is attained through the elimination of the features that produce noise or, are strictly correlated with other already selected features. Our algorithm can improve classification accuracy with a small and appropriate feature subset. Proposed algorithm is easily implemented and because of use of a simple filter in that, its computational complexity is very low. The performance of the proposed algorithm is evaluated through a real Rotary Cement kiln dataset. The results show that our algorithm outperforms existing algorithms.

Keywords—Binary Ant Colony algorithm, Support Vector Machine, feature selection, classification.

I. INTRODUCTION

Our work falls under the Condition monitoring and diagnosis of industrial system which is an important field of engineering study (in our case is a Rotary Cement kiln, see fig. 1). In substance, condition monitoring is a classification problem [12]. The principal function of the condition monitoring is to check the operating condition of the system. It is made up of two parts which are detection and the diagnosis. The phase of detection makes it possible to determine the state of the system as being normal or abnormal. The phase of diagnosis consists in identifying the failing components and to find the causes starting from a whole of symptoms observed [7, 10, 12].

An industrial system is described by a vector of numeric or nominal features. Some of these features may be irrelevant or redundant. Avoiding irrelevant or redundant features is important because they may have a negative effect on the accuracy of the classifier [7, 10]. In addition, by using fewer features we may reduce the cost of acquiring the data and improve the comprehensibility of the classification model (fig. 2).

Feature extraction and subset selection are some frequently used techniques in data pre-processing. Feature extraction is a process that extracts a set of new features from the original features through some functional mapping [15]. Subset selection is different from feature extraction in that no new features will be generated, but only a subset of original features is selected and feature space is reduced [14].

Fig. 1 Rotary Cement kiln

Fig. 2 Construction of the set of features

The idea behind the selection approach is very simple and is shown in Fig. 3. Any method of selection of features consists of four essential points:

A starting subset, which represents the subset of features, initially is used by a procedure of research. This set can be empty, or contains all the features or a random subset. The procedure of research is the essential element of any method of selection. It turns over as result the subset of features which answer the quality standard better. This criterion is turned over by a function of evaluation. This function determines the
quality of classification obtained by using a subset of feature. A criterion of stop is used to finish the procedure of research. This criterion depends to the evaluation function or with the parameters of configuration which are defined by the user [1]. We present in this paper a hybrid approach based on ant colony optimization (ACO) and support vector machine (SVM) for feature selection problems using datasets from the field of industrial diagnosis. This paper presents a novel approach for heuristic value calculation, which will reduce the set of available features. The rest of this paper is organized as follows. In section 2, different methods for feature selection problems are presented. An introduction on ACO applications in feature selection problems is discussed in Section 3. A brief introduction of SVM is presented in Section 4. In Sections 5 and 6, the proposed algorithm is discussed, followed by a discussion on the experimental setup, datasets used and the results.

II. FEATURE SUBSET SELECTION

Feature selection is included in discrete optimization problems. The whole search space for optimization contains all possible subsets of features, meaning that its size is \(2^n\) where \(n\) is the dimensionality (the number of features). Usually FS algorithms involve heuristic or random search strategies in an attempt to avoid this prohibitive complexity. However, the degree of optimality of the final feature subset is often reduced [2, 4, 8].

Two broad categories of optimal feature subset selection have been proposed based on whether feature selection is performed independently of the learning algorithm that constructs the classifier. They are the filter approach and the wrapper approach [3, 22]. The filter approach initially selects important features and then the classifier is used for classification while the wrapper uses the intended learning algorithm itself to evaluate the usefulness of features [13]. The two famous algorithms of this category are Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) [1, 3]. In Sequential forward selection, the features are sequentially added to an empty candidate set until the addition of further features does not decrease the criterion but in Sequential backward selection the features are sequentially removed from a full candidate set until the removal of further features increase the criterion. In our work, we use a hybrid wrapper/filter approach aiming to explore the qualities of both strategies and try to overcome some of their deficiencies [22]. The criterion of stop represents the dimension of the vector obtained by the algorithm where the quality standard does not evolve/move if we add another feature [9].

The first good use of ACO for feature selection seems to be reported in [9]. A. Al-Ani [9] proposes to use a hybrid evaluation measure that is able to estimate the overall performance of subsets as well as the local importance of features. A classification algorithm is used to estimate the performance of subsets. On the other hand, the local importance of a given feature is measured using the Mutual Information Evaluation Function. Susana [19] proposes an algorithm for feature selection based on two cooperative ant colonies, which minimizes two objectives: the number of features and the classification error. The first colony determines the number (cardinality) of features and the second selects the features based on the cardinality given by the first colony. C.L. Huang [20] presents a hybrid ACO-based classifier model that combines ant colony optimization (ACO) and support vector machines (SVM). In his work, an ant’s solution represents a combination of the feature subset and the classifier parameters, and the search of the best solution is performed independently of the learning algorithm that constructs SVM classifier. The classification accuracy and the feature weights of the constructed SVM classifier are used to design the pheromone update strategy. Based on the pheromone table and measured relative feature importance, the transition probability is calculated to select a solution path for an ant. The major inconvenience with this method is the parameters of classifier are fixed during the execution of program and they may have different value in each solution.

III. ANT COLONY OPTIMIZATION (ACO)

Ant colony optimization (ACO) is based on the cooperative behavior of real ant colonies, which are able to find the shortest path from their nest to a food source. ACO algorithms can be applied to any optimization problems that can be characterized as follows [5, 16]:

1. A finite set of components \(C = \{c_1, c_2, \ldots, c_N\}\) is given.
2. A finite set of \(L\) of possible connections/transitions among the elements of \(C\) is defined over a subset \(C'\) of the Cartesian product \(C \times C\), \(L = \{C(C) \mid (c_i, c_j) \in C', |L| \leq N2^c\}\).
3. For each \(C(C) \in L\) a connection cost function \(JC(C) = J(iC(C), t)\), possibly parametrized by some time measure \(t\), is defined.
4. A finite set of constraints \(\Omega = \Omega(C, L, t)\) is assigned over the elements of \(C\) and \(L\).
5. The states of the problem are defined in terms of sequences \(s = (c_1, c_2, \ldots, c_k, \ldots)\) over the elements of \(C\) or of \(L\). \(S\) is a subset of \(S\). The elements in \(S\) define the problem’s feasible states.
6. A neighborhood structure is assigned as follows: the state \(s_2\) is said to be a neighbor of \(s_1\) if \(s_2\) and \(s_1\) are in \(S\) and the state \(S_1\) can be reached from \(s_1\) in one logical step, that is, if \(c_i\) is the last component in the sequence determining the state \(s_1\), it must exists \(c_i, c_j \in C\) such that \(iC(c_j) \in L\) and \(s_2 = (s_1, c_k)\).
7. A solution \(\Psi\) is an element of \(S\) satisfying all the problem’s requirements. A solution is said multi-dimensional if it is defined in terms of multiple distinct sequences over the elements of \(C\).
8. A cost $J_{\phi}(L, t)$ is associated to each solution $\mathcal{V}$, $J_{\phi}(L, t)$ is a function of all the costs $J_{(C)}$ of all the connections belonging to the solution.

It is worth mentioning that ACO makes probabilistic decision in terms of the artificial pheromone trails and the local heuristic information. This allows ACO to explore larger number of solutions than greedy heuristics. Another characteristic of the ACO algorithm is the pheromone trail evaporation, which is a process that leads to decreasing the pheromone trail intensity over time. Pheromone evaporation helps in avoiding rapid convergence of the algorithm towards a sub-optimal region [5, 9, 16].

IV. SUPPORT VECTOR MACHINES

In our wrapper approach, we have used SVM as classifier. SVM is an attractive learning algorithm first introduced by Vapnik [23]. It has a competitive advantage Compared to neural networks, and decision trees.

Given a set of data $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. Where $x_i \in \mathbb{R}^N$ is features vector and $y_i \in \{-1, +1\}$ is a class label. The goal of the SVM is to find an of the form

$$w^T x + b = 0 \quad \text{with} \quad y_i(w^T x_i + b) \geq 1 - \xi_i$$

that separating the S set of training data into two classes (positive and negative) (fig. 5). In general, $S$ cannot be partitioned by a linear hyperplane. However $S$ can be transformed into higher dimensional feature space for making it linearly separable.

![Fig. 5 Two-Class SVM used in linear classification](attachment:image.png)

The mapping $\phi(x)$ need not be computed explicitly; instead, an inner product Kernel of the form

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$ (2)

To solve the optimal hyperplane problem, we can construct a Lagrangian and transforming to the dual. Then, we can equivalently maximize

$$\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

subject to

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \leq C$$

For a test example $z$, we define the decision function as follow

$$\text{sign} (\sum_{i=1}^{m} \alpha_i y_i K(x_i, z) + b)$$

where

- $w$ is the weight vector.
- $b$ is the bias term.
- $C$ is the punishment parameter.
- $\alpha$ is the Lagrange multiplier.

In the next section, we present our proposed SVM/Binary ACO algorithm [6], and explain how it is used for selecting an appropriate subset of features.

V. PROPOSED APPROACH

A. Description of the proposed approach

This research proposes a new implementation of Binary ACO algorithm [6] applied to feature selection, where the best number of features is determined automatically. In this approach, each ant searches the same routine, and pheromone is left on each edge. As an intelligent body, each ant just chooses one edge of the two as shown in Fig 6. The intelligent behavior of ant is very simple, and the incidence matrix traversed by each ant needs only $2 \times n$’s space, which to some extent solves the descriptive difficulty generated from long coding and the reduction of solution quality.

![Fig. 6 The final net obtained by the binary ACO algorithm](attachment:image.png)

B. Probabilistic rule

Initially, the quantity of information in each routine is randomly generated. During the movement, ant $k$ shifts its direction according to the values of pheromone concentration FP and the heuristic value FH. The heuristic value FH is computed using the Fisher discriminant criterion for feature selection [11] [17], which determines the importance of each feature, and it is described in more detail in Section 5.4. The probability that an ant $k$ chooses the feature $X_i$ is given by:

$$P_{S_{ni}} = \frac{FP_i + \frac{FP_i}{\text{Max}(FH)}}{FP_i + \frac{FP_i}{\text{Max}(FH)}}$$ (6)

C. Updating rule

After all ants have completed their solutions, pheromone evaporation on all nodes is triggered, and then according to (2), pheromone concentration in the trails is updated.

$$FP \leftarrow (1 - \rho) FP + \Delta FP$$ (7)

Where $\rho \in [0,1]$ is the pheromone evaporation and $\Delta FP$ is the pheromone deposited on the trails by the ant $k$ that found the best solution for this tour:
\[ \Delta F_P = \frac{1}{1 + F(V') - F(V''')} \] (8)

Where \( F(V') \) represents the best solution built since the beginning of the execution and \( F(V'') \) represents the best solution built during the last tour.

\( F \) is the objective function of our optimization algorithm and \( V \) is the solution funded by the ant \( k \).

The optimal subset is selected according to classifier performance and their length.

The results of this wrapper approach will be compared to a filter approach. The filter approach uses \( F(V') \) an evaluation function. \( F \) is calculated using two concepts: the variance in each class and the variance between classes.

\[ F(V') = \text{trace}\left( \sum_{c=1}^{C} \sum_{i} \left( X_{cv} - m_c \right) \left( X_{cv} - m_c \right)' \right) \] (9)

Where the matrix of variance intra-class is calculated as follows:

\[ \sum_{c=1}^{C} \sum_{i} \left( X_{cv} - m_c \right) \left( X_{cv} - m_c \right)' = \sum_{c=1}^{C} \sum_{i} \left( X_{cv} - m_c \right)' \left( X_{cv} - m_c \right) \] (10)

Whereas the matrix of the variance inter-classes is calculated as follows:

\[ \sum_{c=1}^{C} \sum_{i} \left( m_c - m \right)' \left( m_c - m \right) \] (11)

With:
- \( m \): General centre of gravity
- \( M \): A number of classes
- \( m_c \): Centre of gravity of the class number \( C \)
- \( X_{cv} \): \( V \) th vector of the class number \( C \)
- \( NR_c \): A number of vectors of the class number \( C \)
- \( NR \): Numbers total vectoriel

D. Heuristics

The heuristic value is computed using the Fisher discriminant criterion for feature selection [17]. Considering a classification problem with \( M \) possible classes, the Fisher discriminant criterion is described as follow:

\[ FH(\alpha) = \frac{\sum_{\alpha=1}^{M} \sum_{\alpha=1}^{N} \left( m_{\alpha}(\alpha) - m_{\alpha}(\alpha) \right)}{N, \sigma_\alpha^2(\alpha) - N, \sigma_\alpha^2(\alpha)} \] (12)

Where:
- \( M \) represents the number of classes;
- \( m_{\alpha}(\alpha) \) Represent the centre of gravity of the class number \( C \) by considering only the parameter \( \alpha \) it is calculated as follows:

\[ m_{\alpha}(\alpha) = \frac{1}{N} \sum_{i=1}^{N} X_{\alpha} (\alpha) \] (13)

With \( X_{\alpha} \) is the number \( v \) of the class number \( C \), the value of \( NR \) equal to the number of vectors of the class in question is the vector.

\( \sigma_\alpha^2(\alpha) \) is the variance of the component \( \alpha \) of the vectors of the class number \( C \).

\[ \sigma_\alpha^2(\alpha) = \frac{1}{N} \sum_{i=1}^{N} \left( X_{\alpha}(\alpha) - m_{\alpha}(\alpha) \right)^2 \] (14)

Algorithm 1 presents the description of the Binary ACO-SVM feature selection algorithm.

\begin{algorithm}
\caption{BINARY ACO/SVM FEATURE SELECTION ALGORITHM}
1) \textbf{Initiate the pheromone of the net;}
2) \textbf{Compute the} \( FH(\alpha) \) \textbf{using (7);}
3) \textbf{Ants search using (1);}
4) \textbf{Evaluate the solutions founded by ant colony algorithm using SVM classifier, and reserve the optimal;}
5) \textbf{Upgrade pheromone of the net by optimal solution using (2);}
6) \textbf{Judge whether the stopping condition of the qualification is met, if qualified, ends; otherwise, Goto 3;}
\end{algorithm}

The time complexity of proposed algorithm is \( O(Im) \), where \( I \) is the number of iterations, and \( m \) the number of ants. This can be seen from Fig. 6. In the worst case, each ant selects all the features. As the objective function is evaluated after all ants have completed their solutions, this will result in \( m \) evaluations. After \( I \) iterations, the objective function will be evaluated \( Im \) times.

VI. EXPERIMENTAL RESULTS

A. Data of test

The experimental results comparing the binary ACO algorithm with genetic algorithm are provided for real-life dataset (Vehicle) [18] and industrial dataset (Rotary Cement kiln) [7]. The Vehicle base is collected in 1987 by Siebert. Vehicle consists of 846 recordings which represent 4 classes.
RCK consists of 200 recordings which represent 4 classes.

Fig. 8 Flow Diagram of a Rotary Cement Kiln [7]

B. Parameters of the algorithm of selection

Like any other algorithm, before passing to the phase of selection. Some parameters should be fixed. This problem represents one of the disadvantages of the biomimetic methods. Since the values of the parameters are related to the number of individuals and the distribution of the data on the beach of representation. The following table presents the values of the parameters of our algorithm. These parameters are fixed after the execution of several simulations by using as entered a restricted whole of data.

C. Heuristic factor FH

Heuristic factor FH is taken into account that by the ants which have a behavior related to the probability PS. The ants which have a random behavior are used to discover new spaces of research. The following figures represent the values of heuristic factor FH by using the dataset of the Rotary cement kiln and Vehicle dataset.

According to fig. 9, we notice that the 43rd feature has the greatest value of FH. Consequently it will be present in the final subset.

D. Results

We tested the performances of our algorithm by using the evolutionary method of classification ECMC [21] and the following table shows the quality of classification while using:

a) The best discriminating feature;

b) The best subset of features generated;

c) All features.

The implementation platform was implemented in Matlab 7.9, which is a general mathematical development tool. The Bioinformatics Toolbox functions *svmclassify* and *svmtrain* were used as the SVM classifier. The empirical evaluation was performed using an Intel Pentium Dual Core T4400 2.2GHz with 3GB RAM.

Using the parameters presented in the Table.1, the following results were obtained by taking the best solution after 20 BACOs trials. The Table.1 gives the best solutions obtained for each dataset (Rotary Cement kiln & Vehicle). For the two datasets, the FV of the best solution is indicated with the corresponding Rate of error. We conducted a performance comparison between the proposed wrapper-based ACO (ACO–SVM), the filter-based ACO and the filter-based GA.
Table 2 shows that we obtain an acceptable rate of error with the subset generated by our algorithm. It is also noticed that the value of FV reflects well the quality of classification. The Fig. 9 shows the value of FV obtained by each agent during the last iteration using Rotary Cement kiln dataset.

5th iteration which shows the effectiveness and the speed of our algorithm. The time of convergence of the presented algorithm can be reduced using a lower number of ants. This number is related to the number of features in the dataset.

### Table II

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Features</th>
<th>Vehicle</th>
<th>Rotary Cement kiln</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Error rate</td>
<td>F (V)</td>
<td>Error rate</td>
</tr>
<tr>
<td>Wrapper-based ACO/SVM</td>
<td>One feature 75%</td>
<td>0.03525</td>
<td>46%</td>
</tr>
<tr>
<td></td>
<td>Generated subset 11%</td>
<td>0.4717</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>All features 07%</td>
<td>0.7875</td>
<td>10%</td>
</tr>
<tr>
<td>Filter-based ACO/SVM</td>
<td>One feature 75%</td>
<td>0.03525</td>
<td>46%</td>
</tr>
<tr>
<td></td>
<td>Generated subset 13%</td>
<td>0.6537</td>
<td>15%</td>
</tr>
<tr>
<td></td>
<td>All features 07%</td>
<td>0.7875</td>
<td>10%</td>
</tr>
<tr>
<td>Filter-based GA</td>
<td>One feature 75%</td>
<td>0.03525</td>
<td>46%</td>
</tr>
<tr>
<td></td>
<td>Generated subset 11%</td>
<td>0.7717</td>
<td>15%</td>
</tr>
<tr>
<td></td>
<td>All features 07%</td>
<td>0.7875</td>
<td>10%</td>
</tr>
</tbody>
</table>

Fig. 11 The value of FV obtained by each agent during the last iteration

We notice that after the last iteration, more than 33% of agents find the optimal solution. This is due to the pheromone density which is updated at the end of each iteration.

Fig. 12 The best solution obtained during each iteration

Fig. 13 Best number of features for each iteration in Rotary Cement kiln dataset

Fig. 14 The best solution obtained during each iteration (Vehicle)

Fig. 14 shows that our algorithm discards a bigger percentage of features for the case of Vehicle dataset. However, the selected features are not always the same, once there are features that are weakly relevant and have a similar influence in the classifier.

The results given in figures 11-14 and Tables 2 show that our approach (Wrapper-based ACO-SVM) is very precise. In other word, it gives the optimal solution in compare to those of obtained by the other algorithms. In fact, the results obtained on the Rotary Cement kiln dataset show that our approach converges to the global optimum in all of runs.

### VII. CONCLUSION

In this work, a new approach for selecting best discriminates features subset using Binary ACO algorithm is presented. The ACO is chosen for this study because it is the newest metaheuristic. The goal is to select the best subset that is sufficient to perform a good classification and obtain acceptable rate of error. We have tested the proposed method on two datasets. The experimental results indicate that the proposed Binary ACO algorithm can be applied for larger number of features.

The classifier induced in the experiments was a SVM. This classifier was chosen because it does not suffer from the local minima problem, it has fewer learning parameters to select,
and it produces stable and reproducible results, but our wrapper method can be used with any other supervised classifiers.

In the near future, the performance of the proposed algorithm will be compared with other features selection methods to improve that our algorithm achieving equal or better performance. And we will combine our algorithm with other intelligent classifiers, such as neural networks classifiers.

REFERENCES


