International Journal of Engineering

Journal Homepage: www.ije.ir

Fuzzy-rough Information Gain Ratio Approach to Filter-wrapper Feature Selection

A. Moaref, V. Sattari-Naeini*

Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran

P A P E R I N F O

Paper history:
Received 19 January 2017
Received in revised form 14 March 2017
Accepted 07 July 2017

Keywords:
Feature Selection
Fuzzy Rough Sets
Ant Colony Optimization
Filter-Wrapper Method

A B S T R A C T

Feature selection for various applications has been carried out for many years in many different research areas. However, there is a trade-off between finding feature subsets with minimum length and increasing the classification accuracy. In this paper, a filter-wrapper feature selection approach based on fuzzy-rough gain ratio is proposed to tackle this problem. As a search strategy, a modified Ant Colony Optimization (ACO) algorithm is applied on filter phase. ACO has been approved to be a suitable solution in many difficult problems with graph search space such as feature selection. Choosing minimal data reductions among the subsets of features with first and second maximum accuracies is the main contribution of this work. To verify the efficiency of our approach, experiments are performed on 10 well-known UCI data sets. Analysis of the experimental results demonstrates that the proposed approach is able to satisfy two conflicting constraints of feature selection, increasing the classification accuracy as well as decreasing the length of the reduced subsets of features.

***doi**: 10.5829/ije.2017.30.09c.05*

1. INTRODUCTION

Feature selection is one of the important issues in pattern recognition, machine learning, data mining, bioinformatics, etc. It refers to the process in which the best features, that are more effective in predicting the output, are selected so that its principle objective is improvement in output prediction or accuracy of the classifier [1].

In feature selection problem, the irrelevant and redundant features of the model are eliminated. These unnecessary features not only harden training of the model, but they also reduce the performance and increase the noise. Rough set theory is suitable for removing irrelevant and redundant attributes from a given data set. Because the classical rough sets are unable to deal with real valued noisy features, fuzzy rough sets which are the generalization of classical rough sets and fuzzy sets are used to handle these kinds of problems [2].

Feature selection, as a pre-process task, is done for two purposes: first, in order to select the best features that are more effective in output prediction. In this case, the purpose is improving the accuracy of the classifier [1]. The related works in the literature tend to find subsets that are good in general, but might not be good for a particular classifier. Consequently, it is better to use a range of classifiers to show the utility of the resultant subset. Second, in order to find out information about the features. In this case, we do not particularly care about the resulting classification accuracy, but maximizing the dependency degree is of utmost significance. The contribution of this paper is proposing a filter-wrapper feature selection approach based on fuzzy rough sets using a meta-heuristic search strategy that follows the first aim.

Feature selection methods can be divided into two categories: filter method and wrapper one [2]. Learning algorithm is a part of the wrapper method and classification accuracy is used to select the features. In wrappers, selected subsets of the features are ranked according to the output predictive power of these subsets. The wrapper method can yield high classification accuracy for a particular classifier at the cost of high computational complexity and less generalization of the selected features on other classifiers. The wrapper method consists of two phases: In the first phase, with regard to the accuracy of the learning algorithm (classifier accuracy), the best subset

*Corresponding Author's Email: vsnaeini@uk.ac.ir (V. Sattari-Naeini)

Please cite this article as: A. Moaref, V. Sattari-Naeini, Fuzzy-rough Information Gain Ratio Approach to Filter-wrapper Feature Selection, International Journal of Engineering (IJE), TRANSACTIONS C: Aspects Vol. 30, No. 9, (September 2017) 1326-1333

of features is selected. In the second phase, learning algorithm is trained and tested on the reduced data set.

Filter method is independent of the learning algorithm. In this method, goodness of the features will be evaluated based on their intrinsic characteristics, not solely on their impact on the accuracy of the learning algorithm. For this reason, filter methods are faster than wrapper ones. However, ignoring the performance of the model in feature selection is a shortcoming of filter methods. Two steps are also defined for filter method: at first, features are selected by a mathematical or statistical criterion such as dependency degree, correlation, entropy, etc. In the second step, as it is defined for wrapper method, learning algorithm is trained on the reduced data set and tested.

Wrapper method generally outperforms the filter one in terms of the accuracy of the learning model. Thus, many researchers attempt to speed up the convergence of the wrapper algorithm using a combination of filter and wrapper model; in these approaches, feature selection is a part of learning algorithm, which trains the classifier and chooses a subset of features, simultaneously. For this purpose, the filter/wrapper method is built into the classifier structure [3].

1. 1. Search Strategy and Evaluation Measure

In general, each feature selection method needs to use a suitable search strategy inside the features space. Feature selection is an NP-hard optimization problem, and various algorithms have recently been used to search inside the solution space. Knowing the number of the features, there exist different subsets. By increasing the number of features, exhaustive search is impractical; however, it is useful in small search spaces. Since exploring the whole solution space is almost impossible, using the heuristic algorithms is more practical. Greedy algorithm, as a heuristic one, is remedy for NP-hard problems, but it does not guarantee that the optimal solution is found. Fortunately, meta-heuristic search strategies reach reasonably good solutions and many researchers have used these algorithms for feature selection problem, recently. Some of these algorithms are ACO [4], Genetic Algorithm (GA) [5, 6] and Particle Swarm Optimization (PSO) [7].

Application of meta-heuristic methods in the field of feature selection and many other domains has been widely used [8-12] in order to fuzzy-rough feature selection, clustering the gene expression data sets, Analysis of pre-processing and post-processing methods in classification of medical data, design a cost-sensitive classification system and characterize and cluster Web visitors. As a matter of fact, ACO algorithm is broadly used nowadays by some scholars to find solutions for the feature selection problem [13]. Inspired by the real ants' behavior in finding their closest way between the food and the nest, ACO search method is in fact a meta-heuristic algorithm, attracting the scientists' attention to itself. This algorithm has proved to be useful in solving the graph search space problems as well as resolving the problems concerning Traveling Salesman Problem (TSP) [13], graph coloring [14], scheduling [15], and telecommunication network routing [16].

ACO algorithm is believed to have two main criteria which are effective in guiding the ants to find the most appropriate way in the features space. These criteria are pheromone and heuristic information. Pheromone is deposited on the travelled path by any ant, guiding other ants to find the shortest distance between home and food. The pheromone trails are updated when the ants cross the nodes, resulting in increasing the probability of developing high-quality solutions. One of the most important parts of pheromone updating is pheromone evaporation mechanism that, after some repetitions, causes an increase in the amount of the pheromone on the shorter paths.

In the case of evaluation measure in feature selection, although using a dependency degree measure might be useful to select a subset of features that preserves the meaning of the features and is rarely dependent on the other features, it is not appropriate for real life applications in which the aim is to achieve high classification accuracy [17]. However, there is a tendency for gain criterion to prefer the attribute with more refined partition which encourages offering gain ratio as an improved version of gain, based on fuzzy rough sets. As it would be known, each heuristic algorithm utilizes an evaluation measure. In ACO algorithms, heuristic information criterion could be considered as an evaluation measure. It is worth mentioning that the type of the problem determines the factor in choosing this criterion and directly affects the feature selection. Feature selection is performed by a random function in which the amount of the remaining pheromone on the path is major determinant along with the heuristic information of next paths. In this paper, gain ratio based on fuzzy rough set is suggested as heuristic information to improve ACO algorithm for feature selection problems.

The rest of the paper is organized as follows. In Section 2, information measures in rough and fuzzy-rough set theory is reviewed. A feature selection approach based on fuzzy-rough information gain ratio using ACO search algorithm is proposed in Section 3. In Section 4, experiments and comparisons on several data sets have been discussed. Finally, Section 5 concludes the paper.

2. INFORMATION MEASURES IN FUZZY-ROUGH SET THEORY

The fuzzy equivalence relation is central to fuzzy rough sets [2] and $\tilde{R}$ is a fuzzy equivalence relation, if $\tilde{R}$ satisfies:

- Reflectivity: $\tilde{R}(x,y) = 1, \forall x \in X$;
- Symmetry: $\tilde{R}(x,y) = \hat{\mathrm{R}}(x,y), \forall x,y \in X$;
- Transitivity: $\tilde{\mathrm{R}}(x,y) \geq \min\{\tilde{\mathrm{R}}(x,y), \tilde{\mathrm{R}}(y,z)\}$.

$M(\hat{\mathrm{R}})$ represents a relation matrix for $x_i, x_j \in X$, where R is a fuzzy equivalence relation defined on a nonempty finite set X.

$$M(\tilde{R}) = \begin{bmatrix} r_{1,1} & \cdots & r_{1,n} \\ \vdots & \ddots & \vdots \\ r_{n,1} & \cdots & r_{n,n} \end{bmatrix} \quad (1)$$

Here, $r_{ij} \in [0,1]$ is the relation value of x_i and x_j that can be written as $\tilde{R} = (x,y)$. For the crisp rough set model, $r_{ij} = 1$ if x_i equals to x_j with respect to the crisp equivalence relation R, then $r_{ij} = 1$; otherwise $r_{ij} = 0$. A similarity function that has been used to calculate the equivalence relation is shown by Equation (2), where x_i and x_j are attribute values of two objects on attribute a; a_{max} and a_{min} are maximal and minimal values of attribute a, respectively.

$$r_{ij} = \begin{cases} 1 - 4 \times \frac{|x_i - x_j|}{|a_{max} - a_{min}|} & \frac{|x_i - x_j|}{|a_{max} - a_{min}|} \leq 0.25 \\ 0 & otherwise \end{cases} \quad (2)$$

Two important operations on fuzzy equivalence relations, useful to improve this relation, are defined by:
$\tilde{R} = \widetilde{R_1} \cup \widetilde{R_2} \Leftrightarrow \tilde{R}(x,y) = \max\{\widetilde{R_1}(x,y), \widetilde{R_2}(x,y)\}$
$\tilde{R} = \widetilde{R_1} \cap \widetilde{R_2} \Leftrightarrow \tilde{R}(x,y) = \min\{\widetilde{R_1}(x,y), \widetilde{R_2}(x,y)\}$

Definition 1: The fuzzy partition of the universe U generated by $\tilde{R}$, is defined as [18, 19]:

$$\frac{U}{\tilde{R}} = \{[x_i]_{\tilde{R}}\}_{i=1}^{n} \quad (3)$$

where $\tilde{R}$ is a fuzzy equivalence relation; $[x]_{\tilde{R}}$ is the fuzzy equivalence class equal to $\frac{r_{i1}}{x_1} + \frac{r_{i2}}{x_2} \ldots + \frac{r_{in}}{x_n}$.

Definition 2: The cardinality $[x]_{\tilde{R}}$ is defined as [18, 19]:

$$|[x]_{\tilde{R}}| = \sum_{j=1}^{n} r_{ij} \quad (4)$$

Definition 3: Information quantity of the fuzzy attribute set or the fuzzy equivalence relation is defined as [18, 19]:

$$H(\tilde{R}) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{|[x_i]_{\tilde{R}}|}{n} \quad (5)$$

Definition 4: The joint entropy of B and E is defined as [18, 19]:

$$\tilde{H}(BE) = H(\tilde{R}_B \tilde{R}_E) = -\frac{1}{n}\sum_{i=1}^{n} \frac{|[x_i]_{\tilde{B}} \cap [x_i]_{\tilde{E}}|}{n} \quad (6)$$

where $FIS =< U, A, V, f >$ is a fuzzy information system; A is the attribute set; B and E are two subsets of A.

Definition 5: Let $FIS =< U, A, V, f >$ is a fuzzy decision system, C is the condition attribute set, D is the decision attribute and $B \subseteq C$. Conditional entropy of D conditioned to B is defined as follows, where $[x_i]_{\tilde{R}}$ and $[x_i]_{\tilde{D}}$ are fuzzy equivalence classes containing x_i generated by B and D respectively [15, 20].

$$\tilde{H}(D|B) == -\frac{1}{n}\sum_{i=1}^{n} \frac{|[x_i]_B \cap [x_i]_D|}{|[x_i]_B|} \quad (7)$$

If. $FIS =< U, A, V, f >$ is a fuzzy information system and $B, D \subseteq A$, according to [16, 21, 22] it is known that:

$$\tilde{I}(B;D) = \tilde{H}(D) - \tilde{H}(D|B) \quad (8)$$

$$\tilde{H}(D|B) = \tilde{H}(D|B) - \tilde{H}(B) \quad (9)$$

Definition 6: As a result of Equations (20) and (21), the mutual information of B and D is defined as [23]:

$$\tilde{I}(B;D) = \tilde{H}(D) - \tilde{H}(D|B) = \tilde{H}(D) + \tilde{H}(B) - \tilde{H}(BD) \quad (10)$$

Definition 7: In decision system $FDS =< U, C, \cup, D, V, f >$the gain of attribute a, $\widetilde{Gain}(a,B,D)$, can be defined as [23]:

$$\widetilde{Gain}(a,B,D) = \tilde{I}(B \cup \{a\}; D) - \tilde{I}(B;D) \quad (11)$$

where C is the condition attribute set, D is the decision attribute, and $B \subseteq C$.

Definition 8: Considering definition 13, the mutual information gain ratio of attribute a can be defined as [23]:

$$\widetilde{Gain} - Ratio(a,B,D) = \frac{\widetilde{Gain}(a,B,D)}{\tilde{H}(\{a\})} = \frac{\tilde{I}(B \cup \{a\};D) - \tilde{I}(B;D)}{\tilde{H}(\{a\})} \quad (12)$$

3. A NEW FILTER-WRAPPER FEATURE SELECTION APPROACH

In this section, a new filter-wrapper approach for feature selection in fuzzy-rough sets is described. In this approach, while filter phase utilizes a modified ACO search strategy which is able to do feature selection task as a multi-modal problem, wrapper phase includes a learning model that evaluates the chosen subsets of features and calculates pheromones changes in the selected subsets. Choosing the subsets of features with first and second maximum accuracies as candidate subsets for minimal data reductions is a contributory factor in this work; so each chosen minimal subset has a short length along with an acceptable accuracy value; consequently, the approach is able to satisfy both increase the accuracy and decrease the length of reduced subsets, concurrently.

Figure 1 represents the filter/wrapper method stages for this new fuzzy-rough feature selection approach. Initially, the feature selection problem space is depicted in the form of a complete non-directed graph. The nodes in the graph represent the features and the edges stand for the probability of choosing the next node. In the flowchart of Figure 1, 2nd, 3rd, and 5th stages can be realized easily. The remaining stages are described in the following subsections. In 4th stage in the filter phase, the transition rule, introduced in [14], is used for exploring the nodes space. Node j, as a candidate for

selection, is selected with a probability of 0.5 using Equation (13). If node j is selected, the ant is put on it and node j is removed from the set of available nodes for the ant; otherwise, it is removed from s_k. In order to select another node, Equation (14) is used to calculate the probability of selection of other nodes in s_k where $\eta_j = \widetilde{GainRatio}(j, N_k, D)$ is calculated by Equation (12) as the heuristic information and N_k is regarded as a set of selected nodes by ant k and $\tau_{ij}$is the pheromone value of edge ij.

$$P_{ij}^k = \begin{cases} 1 & j = \text{argmax}_{j \in S_k}\{\tau_{ij}^{\alpha}, \eta_{ij}^{\beta}\} \\ 0 & otherwise \end{cases} \quad (13)$$

$$P_{ij}^k = \begin{cases} \frac{[\tau_{ij}(t)]^{\alpha}.[\tau_{xj}]^{\beta}}{\sum_{x \epsilon S_k}[\tau_{ix}(t)]^{\alpha}.[\eta_x]^{\beta}} & j \epsilon S_k \\ 0 & otherwise \end{cases} \quad (14)$$

In the above relations it is considered that $\alpha = 0.5$ and $\beta = 1$; the initial value of τ_{ij} is 0.1.

The obtained selection probability is deployed in the roulette wheel mechanism to be used as a means of selecting the next node. The selection of a node results in its removal, along with all other previously checked nodes from the first to the current node in the roulette wheel mechanism, from S_k. This task increases the selection probability of higher quality features in the next stages.

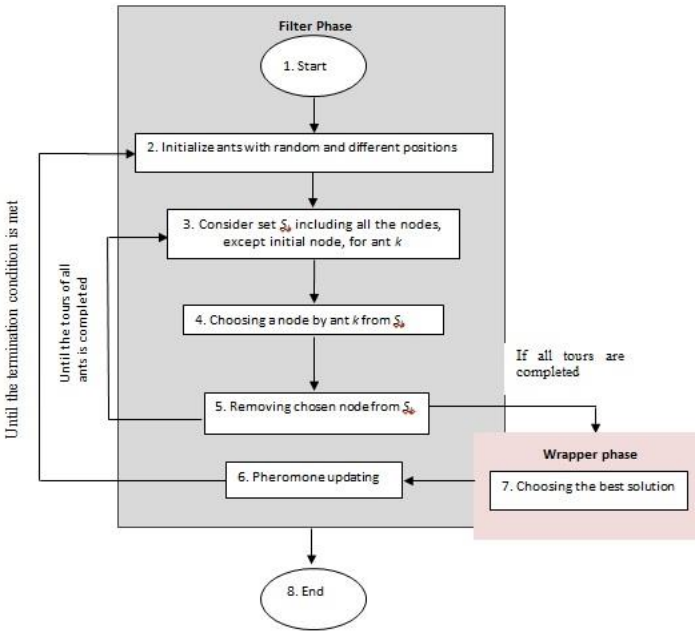

Figure 1. The proposed filter-wrapper method

Consequently, the ants heading for the end of their tour, exploit the space rather than exploring it. After each individual ant creates its own complete tour, the pheromone is updated on the path travelled from the beginning to the end as it has been shown in the 6th stage. As it can be seen in Equation (15), the pheromone evaporates on each edge of the complete graph. Since aim of FS problem (increasing the dependency degree or classification accuracy) relies on this criterion and the main objective of FS is to find fewer features with maximum classification accuracy, the pheromone updating has been performed based on the output of wrapper model according to Equations (16) and (18). In order to maintain the best solutions, observed until the current iteration, the pheromone on the best so far traversed paths is updated based on the output of wrapper model too; See Equations (17) and (18) where $\varphi = 0.5$, $\rho = 0.2$ and BF is the best path traversed in the current iteration and Γ_{Nk} is the accuracy of the classifier as output of the learning model.

$$\tau^{new} = (1 - \rho).\tau^{old} \quad (15)$$

$$\tau_{ij}^{new} = \begin{cases} \tau_{ij}^{old} + \Delta\tau_{ij} & ij \epsilon BF \\ \tau_{ij}^{old}| + \varphi \times \Delta\tau_{ij} & otherwise \end{cases} \quad (16)$$

$$\tau^{new} = \tau^{old} + \varphi \times \Gamma_{Nk} \quad (17)$$

$$\Delta\tau_{ij} = \frac{\Gamma_{Nk}}{length(N_k)} , i, j \epsilon N_k \quad (18)$$

At the end of each iteration, the best observed solutions until now are kept; i.e. in each iteration, the subsets of the features which have maximum accuracy are considered as the best candidate subsets. The subsets

which have the first and the second maximum accuracies among all the best candidate subsets from the first iteration to the current one are kept. Then, the minimal subsets from the kept subsets are considered as the bests in all iterations.

Since wrapper method uses a learning model, feature selection based on wrappers increases the accuracy of the model; however, this method increases the order of mathematical complexity. In this paper, instead of evaluating the features separately, the subsets found by a filter are evaluated using wrapper to decrease the complexity. The output of the wrapper model (accuracy of the classifier) is a criterion for the goodness evaluation of subsets found. After the end of each run, the best seen solution, from the first iteration until the current one, is saved as an optimal solution. In addition to finding out high quality subsets of features, the possibility of finding more than one solution in one run is another advantage of this method compared to other methods that can find only one solution.

4. RESULTS AND DISCUSSION

Eight well-known classifiers are utilized in order to compare the utilization of the methods after feature selection among the proposed method and several other meta-heuristic methods (classical and fuzzy-rough ones). Ant search, Genetic search, and PSO search (utilizing fuzzy-rough dependency degree, CFS Subset Eval and Consistency Subset Eval as heuristic information measures) form nine meta-heuristic methods with which the proposed method is compared to them.

Also, ten UCI repositories of machine learning are utilized for performing experimental results. Since the aim of feature selection problem is often satisfying multiple criteria, such as decreasing the subset cardinality and increasing the model performance (classifier accuracy), a combination of these two criteria is considered, according to Equation (19), as the output results of the methods; here, B is a subset of the features. In this equation, maximization of ψ is caused by both increasing the mean of accuracies and decreasing the length, since the method only considers the subsets of features with first and second maximum accuracies. In this equation, B is a subset of the features. In this equation, maximization of ψ is caused by both increasing the mean of accuracy and decreasing the length, since the method only considers the subsets of features with first and second maximum accuracies.

$$\psi_B = \frac{mean\ of\ accuracies(B)}{length(d)} \quad (19)$$

Table 1 shows the data sets used, and the results are illustrated in Tables 2-11 for these data sets. The left-most columns consist of Subset Evaluator Measures, the 2nd columns are the search algorithms, the 3rd ones the number of minimal reducts obtained by the feature selection methods, the 4th columns the cardinality of the obtained reducts, and the 5th and 6th columns are the mean and variance of accuracy by 8 used classifiers including Part, Nave Bayes, Bayes Net, J48, BFTree, FT, NBTree, Jrip, respectively. Finally, right-most columns, ψ, are calculated using Equation (19).

According to these Tables, this method not only decreases the lengths, but it also increases the level of accuracies in minimal reducts to achieve higher ψ. It means that the lengths of the obtained reducts through the proposed method are either less than other ones or comparable with them. Also, the mean of accuracies in these reducts in the majority of data sets has improved by the method. However, the variance has no significant changes among all the methods.

TABLE 1. Characteristics of used UCI data sets

No.	Data set	No. of features	No. of instances
1	Wine	13	178
2	Pima Indian Diabetes	8	768
3	Glass	9	214
4	Iris	4	150
5	Vote	16	435
6	Parkinsons	22	195
7	Breast Cancer Wisconsin	9	699
8	Sonar	60	208
9	SPECTF Heart	44	80
10	Ecoli	7	336
11	Human Activity Recognition Using Smart Phones	561	10299

TABLE 2. Wine data set results

Subset Eval. Measure	Method	No. of reducts	Size of reducts	Mean of acc.	Var. of acc.	ψ
Proposed method		**3**	4	94.07	3.50	**23.52**
Fuzzy Rough	ACO	1	5	94.10	3.51	18.82
	GA	1	5	89.96	3.36	17.99
	PSO	1	5	89.96	3.36	17.99
CFS	ACO	1	11	94.81	3.59	8.62
	GA	1	11	94.81	3.59	8.62
	PSO	1	11	94.81	3.59	8.62
Consistency	ACO	1	5	94.87	3.60	18.97
	GA	1	5	91.99	3.40	18.40
	PSO	1	5	91.99	3.40	18.40

In addition, the method could find out more than one minimal reduct for Wine, Sonar, and Human Activity Recognition Using Smart Phones data sets as it is illustrated in Tables 2, 8, 9 and 11.

TABLE 3. Pima Indian Diabetes data set results

Subset Eval. Measure	Method	No. of reducts	Size of reducts	Mean of acc.	Var. of acc.	ψ
Proposed method		1	3	74.80	2.60	**24.93**
Fuzzy Rough	ACO	1	8	74.68	2.57	9.33
	GA	1	8	74.68	2.57	9.33
	PSO	1	8	74.68	2.57	9.33
CFS	ACO	1	4	75.03	2.69	18.76
	GA	1	4	75.03	2.69	18.76
	PSO	1	4	75.03	2.69	18.76
Consistency	ACO	1	8	74.68	2.57	9.33
	GA	1	8	74.68	2.57	9.33
	PSO	1	8	74.68	2.57	9.33

TABLE 4. Glass data set results

Subset Eval. Measure	Method	No. of reducts	Size of reducts	Mean of acc.	Var. of acc.	ψ
Proposed method		1	4	63.78	2.07	**15.95**
Fuzzy Rough	ACO	1	8	63.73	2.06	7.97
	GA	1	8	63.73	2.06	7.97
	PSO	1	8	63.73	2.06	7.97
CFS	ACO	1	7	66.06	2.31	9.44
	GA	1	8	67.40	2.39	8.43
	PSO	1	8	67.46	2.40	8.43
Consistency	ACO	1	7	64.84	2.15	9.26
	GA	1	7	64.84	2.15	9.26
	PSO	1	7	64.84	2.15	9.26

TABLE 5. Vote data set results

Subset Eval. Measure	Method	No. of reducts	Size of reducts	Mean of acc.	Var. of acc.	ψ
Proposed method		1	5	95.60	3.65	19.12
Fuzzy Rough	ACO	1	6	94.92	3.54	15.82
	GA	1	11	94.91	3.54	8.63
	PSO	1	12	94.8	3.50	7.90
CFS	ACO	1	6	95.72	3.66	15.95
	GA	1	4	95.63	3.61	**23.90**
	PSO	1	4	95.63	3.61	**23.90**
Consistency	ACO	1	13	94.91	3.52	7.30
	GA	1	11	94.36	3.49	8.58
	PSO	1	10	95.03	3.60	9.50

TABLE 6. Parkinsons data set results

Subset Eval. Measure	Method	No. of reducts	Size of reducts	Mean of acc.	Var. of acc.	ψ
Proposed method		1	4	85.25	3.30	**21.31**
Fuzzy Rough	ACO	1	5	84.93	3.25	16.99
	GA	1	8	83.14	3.09	10.39
	PSO	1	6	84.42	3.15	14.07
CFS	ACO	1	7	84.04	3.11	12.00
	GA	1	10	84.81	3.20	8.48
	PSO	1	10	84.81	3.20	8.48
Consistency	ACO	1	11	84.87	3.21	7.71
	GA	1	8	84.87	3.21	10.61
	PSO	1	9	85.64	3.32	9.52

TABLE 7. Breast Cancer Wisconsin data set results

Subset Eval. Measure	Method	No. of reducts	Size of reducts	Mean of acc.	Var. of acc.	ψ
Proposed method		1	2	94.60	3.69	**47.30**
Fuzzy Rough	ACO	1	7	95.83	3.81	13.69
	GA	1	7	95.69	3.80	13.67
	PSO	1	7	95.78	3.81	13.68
CFS	ACO	1	9	95.62	3.78	10.62
	GA	1	9	95.62	3.78	10.62
	PSO	1	9	95.62	3.78	10.62
Consistency	ACO	1	6	95.55	3.76	15.92
	GA	1	6	95.55	3.76	15.92
	PSO	1	6	95.55	3.76	15.92

TABLE 8. Sonar data set results

Subset Eval. Measure	Method	No. of reducts	Size of reducts	Mean of acc.	Var. of acc.	ψ
Proposed method		**2**	8	73.88	2.49	9.23
Fuzzy Rough	ACO	1	6	68.99	2.10	11.50
	GA	1	9	63.88	2.08	7.10
	PSO	1	6	72.54	2.19	**12.09**
CFS	ACO	1	11	74.34	2.50	6.76
	GA	1	13	71.81	2.18	5.52
	PSO	1	17	75.96	2.52	4.47
Consistency	ACO	1	28	75.30	2.51	2.69
	GA	1	30	73.56	2.48	2.45
	PSO	1	18	71.69	2.17	3.98

TABLE 9. SPECTF Heart data set results

Subset Eval. Measure	Method	No. of reducts	Size of reducts	Mean of acc.	Var. of acc.	ψ
Proposed method		**2**	5	69.37	2.12	**13.87**
Fuzzy Rough	ACO	1	5	72.97	2.20	14.59
	GA	1	8	67.66	2.09	8.46
	PSO	1	5	61.09	2.00	12.22
CFS	ACO	1	6	77.50	2.57	12.92
	GA	1	16	74.47	2.50	4.65
	PSO	1	11	78.28	2.58	7.12
Consistency	ACO	1	19	72.66	2.52	3.82
	GA	1	16	75.66	2.52	4.73
	PSO	1	10	75.00	2.52	7.50

TABLE 10. Ecoli data set results

Subset Eval. Measure	Method	No. of reducts	Size of reducts	Mean of acc.	Var. of acc.	ψ
Proposed method		1	5	81.51	2.95	**16.30**
Fuzzy Rough	ACO	1	6	83.52	3.19	13.92
	GA	1	6	83.52	3.19	13.92
	PSO	1	6	83.52	3.19	13.92
CFS	ACO	1	6	83.52	3.19	13.92
	GA	1	6	83.52	3.19	13.92
	PSO	1	6	83.52	3.19	13.92
Consistency	ACO	1	6	83.52	3.19	13.92
	GA	1	6	83.52	3.19	13.92
	PSO	1	6	83.52	3.19	13.92

TABLE 11. Human Activity Recognition Using Smart Phones data set results

Subset Eval. Measure	Method	No. of reducts	Size of reducts	Mean of acc.	Var. of acc.	ψ
Proposed method		**4**	6	96.83	4.01	**48.28**
Fuzzy Rough	ACO	1	9	97.24	3.98	15.42
	GA	1	9	97.31	3.99	15.42
	PSO	1	9	97.37	4.01	15.40
CFS	ACO	1	16	96.12	3.91	13.61
	GA	1	10	96.12	3.91	13.89
	PSO	1	19	96.12	3.91	13.61
Consistency	ACO	1	32	96.02	3.89	18.12
	GA	1	24	96.02	3.89	18.23
	PSO	1	19	96.02	3.89	18.23

5. CONCLUSION

In this paper a new filter/wrapper approach based on both fuzzy-rough gain ratio and ACO algorithm has been proposed. Its remarkable characteristics are satisfaction of the two important modalities in feature selection simultaneously, i.e. decreasing the length and enhancement of the classification accuracy of the chosen subsets of the features. In addition, our method is able to find several good subsets of features for some data sets. The proposed method has been applied on ten data sets taken from UCI and compared with other meta-heuristic approaches, classical and fuzzy-rough ones, shown in Tables 2-11.

6. REFERENCES

1. Jensen, R. and Shen, Q., "New approaches to fuzzy-rough feature selection", ***IEEE Transactions on Fuzzy Systems***, Vol. 17, No. 4, (2009), 824-838.
2. Liu, H. and Yu, L., "Toward integrating feature selection algorithms for classification and clustering", ***IEEE Transactions on Knowledge and Data Engineering***, Vol. 17, No. 4, (2005), 491-502.
3. Ferreira, A.J. and Figueiredo, M.A., "Efficient feature selection filters for high-dimensional data", ***Pattern Recognition Letters***, Vol. 33, No. 13, (2012), 1794-1804.
4. Jensen, R. and Shen, Q., "Finding rough set reducts with ant colony optimization", in Proceedings of the UK workshop on computational intelligence. Vol. 1, (2003), 15-22.
5. De Stefano, C., Fontanella, F., Marrocco, C. and Di Freca, A.S., "A ga-based feature selection approach with an application to handwritten character recognition", ***Pattern Recognition Letters***, Vol. 35, (2014), 130-141.
6. Zhai, L.-Y., Khoo, L.-P. and Fok, S.-C., "Feature extraction using rough set theory and genetic algorithms—an application for the simplification of product quality evaluation", ***Computers & Industrial Engineering***, Vol. 43, No. 4, (2002), 661-676.
7. Wang, X., Yang, J., Teng, X., Xia, W. and Jensen, R., "Feature selection based on rough sets and particle swarm optimization", ***Pattern Recognition Letters***, Vol. 28, No. 4, (2007), 459-471.
8. Moaref, A. and Naeini, V.S., "A fuzzy-rough approach for finding various minimal data reductions using ant colony optimization", ***Journal of Intelligent & Fuzzy Systems***, Vol. 26, No. 5, (2014), 2505-2513.
9. Mahdizadeh, M. and Eftekhari, M., "Proposing a novel cost sensitive imbalanced classification method based on hybrid of new fuzzy cost assigning approaches, fuzzy clustering and evolutionary algorithms", ***International Journal of Engineering-Transactions B: Applications***, Vol. 28, No. 8, (2015), 1160-1169.
10. Shaeiri, Z. and Ghaderi, R., "Modification of the fast global k-means using a fuzzy relation with application in microarray data analysis", ***International Journal of Engineering-Transactions C: Aspects***, Vol. 25, No. 4, (2012), 283-291.
11. Hamidi, H. and Daraei, A., "Analysis of pre-processing and post-processing methods and using data mining to diagnose heart diseases", ***International Journal of Engineering-Transactions A: Basics***, Vol. 29, No. 7, (2016), 921-930.
12. Hamidzadeh, J., Zabihimayvan, M. and Sadeghi, R., "Detection of web site visitors based on fuzzy rough sets", ***Soft Computing***, (2017), 1-14.

13. Dorigo, M. and Gambardella, L.M., "Ant colony system: A cooperative learning approach to the traveling salesman problem", ***IEEE Transactions on Evolutionary Computation***, Vol. 1, No. 1, (1997), 53-66.
14. Costa, D. and Hertz, A., "Ants can colour graphs", ***Journal of the Operational Research Society***, Vol. 48, No. 3, (1997), 295-305.
15. Merkle, D. and Middendorf*, M., "On solving permutation scheduling problems with ant colony optimization", ***International Journal of Systems Science***, Vol. 36, No. 5, (2005), 255-266.
16. Okdem, S. and Karaboga, D., "Routing in wireless sensor networks using ant colony optimization", in Adaptive Hardware and Systems. AHS. First NASA/ESA Conference on, IEEE., (2006), 401-404.
17. Maji, P. and Garai, P., "On fuzzy-rough attribute selection: Criteria of max-dependency, max-relevance, min-redundancy, and max-significance", ***Applied Soft Computing***, Vol. 13, No. 9, (2013), 3968-3980.
18. Hu, Q., Yu, D., Xie, Z. and Liu, J., "Fuzzy probabilistic approximation spaces and their information measures", ***IEEE Transactions on Fuzzy Systems***, Vol. 14, No. 2, (2006), 191-201.
19. Hu, Q., Yu, D. and Xie, Z., "Information-preserving hybrid data reduction based on fuzzy-rough techniques", ***Pattern Recognition Letters***, Vol. 27, No. 5, (2006), 414-423.
20. Lee, T.T., "An information-theoretic analysis of relational databases—part i: Data dependencies and information metric", ***IEEE Transactions on Software Engineering***, Vol., No. 10, (1987), 1049-1061.
21. Pawlak, Z., "Rough sets: Theoretical aspects of reasoning about data, Springer Science & Business Media, Vol. 9, (2012).
22. Li, J., Mei, C. and Lv, Y., "A heuristic knowledge-reduction method for decision formal contexts", ***Computers & Mathematics with Applications***, Vol. 61, No. 4, (2011), 1096-1106.
23. Dai, J. and Xu, Q., "Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification", ***Applied Soft Computing***, Vol. 13, No. 1, (2013), 211-221.

Fuzzy-rough Information Gain Ratio Approach to Filter-wrapper Feature Selection

A. Moaref, V. Sattari-Naeini

Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran

P A P E R I N F O

Paper history:
Received 19 January 2017
Received in revised form 14 March 2017
Accepted 07 July 2017

Keywords:
Feature Selection
Fuzzy Rough Sets
Ant Colony Optimization
Filter-Wrapper Method

چکیده

سالهاست که مسأله‌ی انتخاب ویژگی در عرصه‌های تحقیقاتی مختلف و برای کاربردهای مختلف به کار می رود؛ درحالی‌که همواره بین پیداکردن مجموعه ویژگی‌ها با کمترین طول از یک طرف، و افزایش دقت دسته‌بند از طرف دیگر تضاد وجود دارد. در این مقاله یک روش فیلتر–دسته بند بر پایه‌ی نرخ بهره‌ی اطلاعاتی در مجموعه‌های ناهموار فازی ارائه شده است که می‌تواند از عهده‌ی این مشکل برآید. از آن‌جاکه الگوریتم بهینه‌سازی کولونی مورچگان (ACO) می‌تواند پاسخ مناسبی برای جستجو در مسایل با فضای گراف از جمله انتخاب ویژگی باشد، در این کار یک الگوریتم تغییر یافته‌ی ACO در فاز فیلتر و به عنوان استراتژی جستجو معرفی شده است؛ اما اصلی‌ترین نوآوری این کار را می‌توان انتخاب مجموعه‌های مینیمم کاهش یافته‌ی ویژگی با اولین و دومین بهترین دقت دسته‌بند در نظر گرفت. ما برای تعیین کارآمدی روش ارایه شده، آن را بر روی ده مجموعه داده‌ی شناخته شده UCI آزمودیم. تحلیل نتایج به دست آمده حاکی از آن است که علی‌رغم روش‌های موجود، روش پیشنهادی ما قادر است هم‌زمان دو شرط متضاد انتخاب ویژگی، یعنی افزایش دقت دسته بند و کاهش طول زیرمجموعه‌های ویژگی را به دنبال داشته باشد.

***doi**: 10.5829/ije.2017.30.09c.05*