Fuzzy-Rough Sample Selection

Javad Rahimipour Anaraki[a] and Chang Wook Ahn[b]

Department of Computer Engineering, Sungkyunkwan University, Suwon 440-746, Korea

[a] j.r.anaraki@ieee.org, [b] cwan@skku.edu

Keywords: Fuzzy-rough set, sample selection, shuffled frog leaping algorithm

Abstract. Rough set theory (RST) is one of the most successful mathematical tools for modeling uncertainty and vagueness. During recent years, many feature selection methods have been proposed based on RST to deal with discrete datasets. As an extension, fuzzy-rough set was introduced to evade the lack of RST in dealing with continuous data in feature selection. However, few investigations were conducted in fuzzy-rough sample selection (FRSS). This paper proposes a new FRSS based on fuzzy-rough positive region (FRPR) as evaluation measure and shuffled frog leaping algorithm (SFLA) as search algorithm. The effectiveness of the proposed method is demonstrated using resulting accuracies of nine classifiers over fifteen UCI datasets. All experimental results show a meaningful increase in classification accuracy and decrease and dataset size, respectively.

Introduction

In nature, selection is the process in which the most effective and powerful objects are selected referring to a measure called fitness. However, this process is used in different aspects of computer science, such as machine learning, pattern recognition and data mining, i.e. feature, sample selection and feature-sample selection. Highly dimensional datasets are generated on daily, hourly and even worse secondly basis. The needed amount of processing power of these much of the data is usually beyond the computational power of existing high-end hardware facilities. Therefore, software methods such as feature and sample selection are introduced to decrease the size of datasets, increase classification accuracy and also overcome the inadequate processing power of current hardware. These methods minimize the effect of noise by removing redundant and irrelevant data.

Feature selection (FS) is one of the most capable machine learning methods to minimize the size of huge datasets and acts as a pre-process to simplify and smoothen the task of the main process (i.e. Classification). In this method, redundant and irrelevant features are filtered out and the most informative features remain untouched. Every selection method needs 1- search algorithm and 2- evaluation measure that the former finds the minimal subsets and the later evaluate the effectiveness of the selection. Sample selection (SS) is also a process which selects highly informative samples using aforementioned elements in FS.

Selecting M features out of N features by means of a comprehensive search is an NP-hard problem. What is worse, it has been proven that approximating the minimal relevant subset is hard up to very large factors [1]. Therefore, greedy search methods and metaheuristic search strategies are suitable for solving this problem. However, all of the greedy search methods suffer from the deficiency of becoming trapped in local optima [2]. Forward and backward search mechanisms are instances of greedy search algorithms that are widely used for FS because of their ideal time complexity; therefore, they are not capable of avoiding local optima [2][3]. Due to this deficiency, metaheuristic search strategies have been widely utilized to solve FS problems [2][4][5][6].

The rough set theory (RST [7]) is one the most successful tools to deal with imperfect knowledge. During recent years, this theory has been applied to different domains. FS based on RST received much of interest due to its capability of confronting discrete data with no human provided information. Therefore, in order to add the capability of dealing with continuous data to the rough set, fuzzy-rough set has been introduced [8]. Based on this combination many feature selection methods were proposed as presented in [9]. However, referring to author's information, only one research has been done in fuzzy-rough instance selection [10], so far. In this article, instances are removed until no uncertainty remains, which could affect positive region.

In this paper, we propose an approach to fuzzy-rough sample selection (FRSS) based on shuffled frog leaping algorithm (SFLA [13]).

Preliminaries

Rough sets. A dataset can be presented as a table where each row shows an object and each column usually is named a feature. Table 1 shows a table in which {Age, LEMS} are conditional attributes and {Walk} is called class or decision attribute. This table also contains seven rows that are named by $x_1, x_2, ..., x_7$, respectively, and are the samples in this table. Also the table is called an Information System and can be presented by pair (U, A), where U is a nonempty finite set of objects called the universe and A is a nonempty finite set of attributes such that $a : u \to V_a$ for every $a \in A$. V_a is the set of values that attribute a may take.

Table 1: An example of decision table

Object	Age	LEMS	Walk
x_1	16-30	50	Yes
x_2	16-30	0	No
x_3	31-45	1-25	No
x_4	31-45	1-25	Yes
x_5	46-60	26-49	No
x_6	16-30	26-49	Yes
x_7	46-60	26-49	No

In this example objects x_3 and x_4 are exactly the same with respect to conditional attributes values. These objects are called *indiscernible* with any subset of P of $A, (P \subseteq A)$. There is an associated equivalence relation.

$$IND(P) = \{(x, x') \in U^2 / \forall a \in P, a(x) = a(x')\}, \tag{1}$$

where $IND(P)$ is called the P-indiscernibility relation. The partition of U produced by $IND(P)$ is denoted $U / IND(P)$ (or for simplicity U / P). The method of calculating such partition has been given in [8].

Let X be a subset of U, approximating subset X using rough set theory is done by means of upper and lower approximation. Upper approximation of X, $(\overline{P}X)$ contains objects which are possibly classified in X, and objects in lower approximation $(\underline{P}X)$ are the ones which are surely classified in X. Boundary region of X can be determined by subtracting upper approximation from lower approximation and where it is a non-empty set, X is called a rough set otherwise it is a crisp set. Rough set is shown by ordered pairs $(\overline{P}X, \underline{P}X)$. Let P and Q be subset of attributes. Different regions are defined using this pair as below:

$$POS_P(Q) = \bigcup_{X \in U/Q} \underline{P}X. \tag{2}$$

$$NEG_P(Q) = U - \bigcup_{X \in U/Q} \overline{P}X. \tag{3}$$

$$BND_P(Q) = \bigcup_{X \in U/Q} \overline{P}X - \bigcup_{X \in U/Q} \underline{P}X. \tag{4}$$

Rough set positive region (RPR) of partition U/Q (denoted by $POS_{\mathrm{P}}(Q)$) is a set of all objects which can uniquely classify to blocks of partition U/Q by means of P [11]. Negative region (denoted by $NEG_{\mathrm{P}}(Q)$) is a set of objects which cannot be classified to the partition U/Q. The boundary region (denoted by $BND_{\mathrm{P}}(Q)$) is the set of objects that can possibly, but not certainly be classified in this way [8].

A set of attributes Q depends totally on a set of attributes P, denoted by $P \underset{k}{\Rightarrow} Q$, if all attribute values from Q are uniquely determined by values of attributes from P. So Q depends on P in degree of k and it is denoted by:

$$k = \gamma_P(Q) = \frac{|POS_{\mathrm{P}}(Q)|}{|U|}. \tag{5}$$

The Equation (5) is the definition of dependency degree.

Based on Equation (5), the QuickReduct algorithm which is given is Fig. 1 calculates a reduct without finding all subsets. It starts from an empty set and each time selects a feature which causes the greatest increase in dependency degree. The algorithm stops when adding more features does not increase dependency degree. It does not guarantee to find minimal reduct as long as it employs a greedy algorithm which is a forward search and capable of being trapped in local optimum.

Fig. 1 QuickReduct Algorithm

```
QUICKREDUCT (C,D)
C, the set of all conditional attributes;
D, the set of decision attributes.
R ← {}
do
    T ← R
    foreach x ∈ (C − R)
        if γ_(R∪{x})(D) > γ_T(D)
            T ← R ∪ {x}
            R ← T
until γ_R(D) == γ_C(D)
return R
```

Fuzzy-rough set. In many cases we face a mixture of crisp and continuous data in datasets that cannot be handled by rough set. The need for a method based on RST for tackling this issue ends to combining fuzzy and rough set theories. Both theories deal with the information granulation problem but with different means. Fuzzy set deals with fuzzy information granulation whereas rough set is concentrated on crisp information granulation [8][14][15]. One way to handle continuous data using rough set is to discretize continuous data in advance and make a new crisp valued dataset. Indeed, discretization is not enough as long as similarity between two values is still unspecified.

Original definitions of *X*-lower and *X*-upper approximations are [8]:

$$\mu_{\underline{P}X}(F_i) = \inf_x \max\left\{1 - \mu_{F_i}(x), \mu_X(x)\right\} \forall i, \tag{6}$$

$$\mu_{\overline{P}X}(F_i) = \sup_x \min\left\{\mu_{F_i}(x), \mu_X(x)\right\} \forall i, \tag{7}$$

F_i is fuzzy equivalence class and $\mu_{F_i}(x)$ is membership degree of object x to a fuzzy equivalence class F_i. The tuple $< \mu_{\underline{X}}, \mu_{\overline{X}} >$ is called fuzzy-rough set. As the memberships of objects are not explicitly available in above mentioned terms, the fuzzy lower and upper approximations are redefined as [8]:

$$\mu_{\underline{P}X}(x) = \sup_{F \in U/P} \min\left\{\mu_F(x), \inf_{y \in U} \max\{1-\mu_F(y), \mu_X(y)\}\right\}, \tag{8}$$

$$\mu_{\overline{P}X}(x) = \sup_{F \in U/P} \min\left\{\mu_F(x), \sup_{y \in U} \min\{\mu_F(y), \mu_X(y)\}\right\}, \tag{9}$$

where F is fuzzy equivalence class, $\mu_F(x)$ is membership degree of object x to fuzzy equivalence class F, $\mu_F(y)$ is membership degree of object y to fuzzy equivalence class F and $\mu_X(x)$ is membership degree of object x to fuzzy equivalence class X.

Proposed method

Our proposed method is composed of two parts, 1- search algorithm and 2- evaluation measure which are explained in following subsections.

Search algorithm. Shuffled Frog Leaping Algorithm (SFLA [13]) is a metaheuristic search algorithm which is inspired by real frogs. The search starts by generating a population over search space. Then the population is divided into sub-population called memeplexes which are able to evolve separately.

In each memeplexes, frogs participates in meme evolution due to infection by other frogs. By meme evolution, each frog performance is increased referring to the best frog in each memeplex and poor ideas evolve toward new ideas. The frogs are infected both by best frogs in their memeplex and the entire population. After specified number of evolution, memeplexes are mixed together and new memeplexes are emerged by shuffling the population. This process migrates frogs to different regions of the swamp. Therefore, they can share their experiences with other frogs.

As presented in [13], SFLA parameter selection should be done based on the properties of the problem; however, it is still untouched for SS. By referring to authors' recommendation, for problems with 15-20 variables, ranges in Table 2 are suggested.

As the number of samples might increase beyond 20 for different datasets, the values for following variables would increase respectively. Therefore, we use the parameters in Table 2 as a reference and recalculate their values as the number of samples increases.

Table 2: SFLA Parameters for FRSS

m	***n***	***N***	***q***	S_{max}
$100 \le m \le 150$	$30 \le n \le 100$	$20 \le N \le 30$	20	100%

where m is the number of memeplexes, n is number of frogs, N is the number of iterations of the evolution process, q is number of randomly selected frogs to form memeplex and S_{max} is the maximum step size allowed after infection.

Evaluation measure. Prior to definition of evaluation measure that is fuzzy-rough positive region (FRPR), the final definitions of X-lower, X-upper approximation and the degree of fuzzy similarity [8] are given in Equations (10) to (12), respectively.

$$\mu_{\underline{R_P}X}(x) = \inf_{y \in \mathbb{U}} I\left\{\mu_{R_P}(x,y), \mu_X(y)\right\}, \tag{10}$$

$$\mu_{\overline{R_P}X}(x) = \sup_{y \in \mathbb{U}} T\left\{\mu_{R_P}(x,y), \mu_X(y)\right\}, \tag{11}$$

$$\mu_{R_P}(x,y) = \bigcap_{a \in P}\left\{\mu_{R_a}(x,y)\right\}, \tag{12}$$

where I is Łukasiewicz fuzzy implicator which is defined by $min(1\text{-}x+y,1)$ and T is Łukasiewicz fuzzy t-norm which is defined by $max(x+y\text{-}1,0)$. Here, R_P is the fuzzy similarity relation and $\mu_{R_P}(x,y)$ is the degree of similarity between object x and y considering feature a. In [15], three classes of Fuzzy-Rough set based on three different classes of implicators, namely S-, R-, and QL-implicators, and their properties have been investigated. One of the best Fuzzy similarity relations as suggested in [8] is given in Equation (14).

$$\mu_{R_a}(x,y)=\max\left(\min\left(\frac{\left(a(y)-\left(a(x)-\sigma_a\right)\right)}{\left(a(x)-\left(a(x)-\sigma_a\right)\right)},\frac{\left(\left(a(x)+\sigma_a\right)-a(y)\right)}{\left(\left(a(x)+\sigma_a\right)-a(x)\right)}\right),0\right), \tag{13}$$

where $a(x)$ and $a(y)$ are values of objects x and y referring to feature a. The σ_a is variance of feature a.

Fuzzy-rough sample selection can be conducted for real-valued datasets using the lower approximation. The RPR is defined as a union of lower approximations and by referring to the extension principle [8], the membership of object x to a FRPR is given in Equation (14).

$$\mu_{POS_{R_P}(Q)}(x)=\sup_{X\in \mathbb{U}/Q}\mu_{\underline{R_P}X}(x), \tag{14}$$

where $\mu_{\underline{R_P}X}(x)$ is lower approximation as defined in Equation (10). If the equivalence class of which x belongs to, does not belong to positive region, obviously x won't be a part of positive region.

The proposed FRSS is based on FRPR as evaluation measure and SFLA as a search method. In each iteration the SFLA selects a subset of samples based on the value of FRPR. The length of each frog in the population is equal to the number of samples in the dataset where their presence and absence are depicted by one and zero, respectively. Each frog's formation is shown in Fig. 2.

Fig 2. Each Frog's Formation in FRSS

1	…	0

As SFLA generates initial population, related dataset formations are constructed referring to each frog's individual. Based on the Table 1, a possible frog's formation and related dataset are presented in Fig. 3 and Table 3, respectively.

Fig 3. A Possible Frog's Formation in FRSS

1	0	1	1	0	1	0

Table 3: Resulting Dataset Referring to Possible Frog's Formation

Object	Age	LEMS	Walk
x_1	16-30	50	Yes
x_3	31-45	1-25	No
x_4	31-45	1-25	Yes
x_6	16-30	26-49	Yes

As shown in Fig. 3, the 1st, 3rd, 4th, 6th positions in frog's formation are equal to one which means these samples are selected and the rest has to be removed. Therefore, referring to this information, the resulting dataset is shown in Table 3.

Then, fitness of all frogs (in other words resulting datasets) is calculated using FRPR as shown in Equation 14 by considering the whole features and selected samples. This process continues until either highest value of FRPR or the maximum number of iterations is reached.

Experimental results

Fifteen datasets were taken from the UCI depository of machine learning [16] to perform experiments. These datasets are selected from different varieties with a wide range of number of samples. The characteristics of these datasets are shown in Table 4.

In Table 5 number of selected samples are presented and compared with unreducted datasets. The last row shows the mean of the number of samples in unreducted datasets and reducted ones. By using the proposed method, all datasets experienced 52.65% decrease in the number of samples in average which causes considerable saving in both memory and processing power.

As long as FRSS acts as a pre-process, so the main process which is usually classification can be done in a more efficient way by means of time and space complexity.

After SS by proposed method, nine classifiers such as PART, JRip, Naive Bayes, Bayes Net, J48, BFTree, FT, NBTree and RBFNetwork have been employed to classify the results based on 10-fold cross validation. All classifiers have been implemented efficiently in Weka that is machine learning tool [17].

The mean accuracies of all classifiers are presented in Table 6 for both unreducted and reducted datastets. The last row of this table shows the mean of the mean of classification accuracies. For eleven datasets out of fifteen, FRSS causes an increase in classification accuracies whereas for Heart, Ionoshpere, Olitos and Soybean couldn't improve the classification accuracies. The Wine dataset has experienced the highest increase in classification accuracy 10.18% and the Soybean dataset has experienced the highest negative impact of -8.04%.

The FRSS cause 1.97% increase in classification accuracy in average. It is concluded that FRSS is suitable in simplifying the classification process by selection most informative samples and leading to less memory and computational power usage.

Table 4: Datasets characteristics

Datasets	**Samples**	**Features**
Blood Transfusion	748	4
Breast Cancer	683	9
Breast Tissue	106	9
Cleveland	297	13
Glass	214	9
Heart	270	13
Ionosphere	351	33
Lung Cancer	27	56
Olitos	120	25
Parkinson	195	22
Pima Indian Diabetes	768	8
Sonar	208	60
Soybean	47	35
SPECTF Heart	80	44
Wine	178	13

Table 5: Number of selected Samples by FRSS

Datasets	**Unreducted**	**FRSS**
Blood Transfusion	748	264
Breast Cancer	683	256
Breast Tissue	106	70
Cleveland	297	199
Glass	214	144
Heart	270	156
Ionosphere	351	115

Lung Cancer	27	20
Olitos	120	81
Parkinson	195	130
Pima Indian Diabetes	768	256
Sonar	208	140
Soybean	47	31
SPECTF Heart	80	55
Wine	178	115
Mean	286.13	135.47

Table 6: Mean of classification accuracies (%)

Datasets	**Unreducted**	**FRSS**
Blood Transfusion	77.20	77.30
Breast Cancer	96.18	96.40
Breast Tissue	66.46	68.66
Cleveland	50.13	50.88
Glass	61.89	66.87
Heart	79.55	73.61
Ionosphere	89.68	89.55
Lung Cancer	55.56	57.61
Olitos	69.81	69.07
Parkinson	82.34	85.64
Pima Indian Diabetes	75.00	75.61
Sonar	67.47	74.73
Soybean	98.58	90.54
SPECTF Heart	73.06	73.74
Wine	85.52	95.70
Mean	75.23	77.20

Conclusion

In this paper a new fuzzy-rough sample selection (FRSS) based on shuffled frog leaping algorithm (SFLA) has been proposed. The performance of the proposed method by referring to the number of selected samples and classification accuracies resulting from nine classifiers were compared with unreducted datasets. The proposed FRSS cause 52.65% decrease in memory usage and 1.97% increase in classification accuracy in average. For eleven datasets out of fifteen, FRSS causes an increase in classification accuracies whereas for Heart, Ionoshpere, Olitos and Soybean couldn’t improve the classification accuracies. As a future work, FRSS can be combined with different metaheuristic methods such as GA, PSO and ACO.

Acknowledgment

This work was supported by the NRF grant funded by the Korean government (MSIP) (NRF-2012R1A2A2A01013735).

References

[1] Nock, R., Sebban, M., (2000) Sharper bounds for the hardness of prototype and feature selection. In: Arimura H, Jain S, Sharma A, editors, Algorithmic Learning Theory, Springer Berlin Heidelberg, volume 1968 of Lecture Notes in Computer Science. pp. 224-238.

[2] Yusta, S. C., (2009) Different metaheuristic strategies to solve the feature selection problem. Pattern Recognition Letters 30: 525 - 534.
[3] Pudil P., Novoviov J., Somol, P., (2002) Feature selection toolbox software package. Pattern Recognition Letters 23: 487 - 492.
[4] ElAlami, M., (2009) Alter model for feature subset selection based on genetic algorithm. Knowledge-Based Systems 22: 356 - 362.
[5] Nemati, S., Basiri, M. E., Ghasem-Aghaee, N., Aghdam, M. H., (2009) A novel ACO-GA hybrid algorithm for feature selection in protein function prediction. Expert Systems with Applications 36: 12086 -12094.
[6] Vieira, S. M., Sousa, J. M., Runkler, T. A., (2010) Two cooperative ant colonies for feature selection using fuzzy models. Expert Systems with Applications 37: 2714 - 2723.
[7] Pawlak, Z., Rough Sets. Intern. J of Comput. and Inf. Sci., 11(1982) 341-356.
[8] Jensen, R. & Shen, Q., New Approaches to Fuzzy-Rough Feature Selection. IEEE Trans. On Fuzzy Syst. 17-4 (2009) 824-838.
[9] Anaraki, J. R. & Eftekhari, M., Rough set based feature selection: A review, in Information and Knowledge Technology (IKT), 2013 5th Conference on, May 2013, pp. 301-306.
[10] Jensen, R. and Cornelis, C. Fuzzy-rough instance selection. International Conference on Fuzzy Systems (FUZZ), 1776-1782, Barcelona (2010).
[11] Komorowski, J., Pawlak, Z., Polkowski, L., & Skowron, A. (1999). Rough Sets: A Tutorial. Rough-Fuzzy Hybridization: A New Trend in Decision Making, 3-98.
[12] Shen, Q., & Jensen, R. (2004). Selecting informative features with Fuzzy-Rough sets and its application for complex systems monitoring. Pattern Recognition, 37, 1351–1363.
[13] Eusuff, M., Lansey, K. and Pasha, F., “Shuffled frog-leaping algorithm: a memetic metaheuristic for discrete optimization,” Engineering Optimization, vol. 38, no. 2, pp. 129-154, 2006
[14] Degang, C., Suyun, Z. Local reduction of decision system with FuzzyRough Sets, Fuzzy Sets and Systems 161 (2010) 18711883.
[15] Radzikowska, A., Kerre, E. A comparative study on Fuzzy-Rough Sets. Fuzzy Sets Systems 126 (2002) 137155.
[16] Bache, K., and Lichman, M., “UCI machine learning repository,” 2013. [Online]. Available: http://archive.ics.uci.edu/ml
[17] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., et al. (2009) The weka data mining software: An update. SIGKDD Explor Newsl 11: 10-18.