Measures for Unsupervised Fuzzy-Rough Feature Selection

Neil Mac Parthaláin and Richard Jensen
Department of Computer Science,
Aberystwyth University, Wales, UK
{ncm,rkj}@aber.ac.uk

Abstract

For supervised learning, feature selection algorithms attempt to maximise a given function of predictive accuracy. This function usually considers the ability of feature vectors to reflect decision class labels. It is therefore intuitive to retain only those features that are related to or lead to these decision classes. However, in unsupervised learning, decision class labels are not provided, which poses questions such as; which features should be retained? and, why not use all of the information? The problem is that not all features are important. Some of the features may be redundant, and others may be irrelevant and noisy. In this paper, some new fuzzy-rough set-based approaches to unsupervised feature selection are proposed. These approaches require no thresholding or domain information, and result in a significant reduction in dimensionality whilst retaining the semantics of the data.

1. Introduction

Large dimensionality presents a problem for handling data due to the fact that the complexity of many commonly used operations are highly dependent (e.g. exponentially) on the level of dimensionality. The problems associated with such large dimensionality mean that any attempt to use machine learning or data-mining tools to extract knowledge, results in very poor performance. Feature selection (FS) [4] is a process which attempts to select features which are information-rich whilst retaining the original meaning of the features following reduction. Most learning algorithms are unable to consider problems of such size, whilst those that are not will usually perform poorly.

Rough set theory (RST) [12] is an approach that can be used for dimensionality reduction, whilst simultaneously preserving the semantics of the features. Also, as RST operates only on the data and does not require any external information, it is completely data-driven. RST however has one main disadvantage: its inability to deal with real-valued data. In order to tackle this problem, methods of discretising the data were employed prior to the application of RST. The use of such methods can result in information loss however, and a number of extensions to RST have emerged [6, 17] which have attempted to address this inability to operate on real-valued domains. One such approach is fuzzy-rough sets (FRS) which have the ability to operate effectively on real-valued (and crisp) data, thus minimising information loss [8].

Conventional supervised FS methods evaluate various feature subsets using an evaluation function or metric to select only those features which are related to, or lead to, the decision classes of the data under consideration. However, for many data mining applications, decision class labels are often unknown or incomplete, thus indicating the significance of unsupervised feature selection. In a broad sense, two different types of approach to unsupervised FS have been adopted: Those which maximise clustering performance using an index function [5], [11], and those which consider features for selection on the basis of dependency or relevance. The central idea behind the latter, is that any single feature which carries little or no further information than that subsumed by the remaining features is redundant and can therefore be eliminated [7, 10]. The approach described in this paper is related to these techniques since it involves the removal of features which are considered to be redundant.

The work presented here is based on FRS, employing novel fuzzy-rough measures to examine the level of dependency between subsets of features. The remainder of this paper is structured as follows. Section 2 introduces the theoretical background to RST and FRS and their application to FS. Section 3 presents the new unsupervised fuzzy-rough feature selection method. The proposed approach is compared with an advanced supervised FS technique [9] which is also based on FRS, and results are presented in Section 4. The paper is then concluded in Section 5.

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.45

IEEE computer society

2. Supervised rough approaches

There has been great interest in developing methodologies which are capable of dealing with imprecision and uncertainty. The success of rough set theory for this is due in part to the fact that it operates only on the data and does not require any external information. As RST handles only one type of imperfection found in data, it is complementary to other concepts for the purpose, such as fuzzy set theory. The two fields may be considered analogous in the sense that both can tolerate inconsistency and uncertainty - the difference being the type of uncertainty and their approach to it; fuzzy sets are concerned with vagueness, rough sets are concerned with indiscernibility.

2.1. Rough set feature selection

Let $I = (\mathbb{U}, \mathbb{A})$ be an information system, where $\mathbb{U}$ is a non-empty set of finite objects (the universe of discourse) and $\mathbb{A}$ is a non-empty finite set of attributes such that $a : \mathbb{U} \rightarrow V_a$ for every $a \in \mathbb{A}$. V_a is the set of values that attribute a may take. With any $P \subseteq \mathbb{A}$ there is an associated equivalence relation $IND(P)$:

$$IND(P) = \{(x, y) \in \mathbb{U}^2 | \forall a \in P,\ a(x) = a(y)\} \quad (1)$$

The partition of $\mathbb{U}$, generated by *IND(P)* is denoted $\mathbb{U}$*/IND(P)* (or $\mathbb{U}$*/P* for simplicity) and can be calculated as follows:

$$\mathbb{U}/IND(P) = \otimes\{\mathbb{U}/IND(\{a\}) | a \in P\}, \quad (2)$$

where $\otimes$ is specifically defined as follows for sets A and B:

$$A \otimes B = \{X \cap Y | X \in A, Y \in B, X \cap Y \neq \emptyset\} \quad (3)$$

If $(x, y) \in IND(P)$, then x and y are indiscernible by attributes from P. The equivalence classes of the P-indiscernibility relation are denoted $[x]_P$.

Let $X \subseteq \mathbb{U}$. X can be approximated using only the information contained within P by constructing the *P-lower* and *P-upper* approximations of X:

$$\begin{aligned} \underline{P}X &= \{x \in \mathbb{U} | [x]_P \subseteq X\} && (4) \\ \overline{P}X &= \{x \in \mathbb{U} | [x]_P \cap X \neq \emptyset\} && (5) \end{aligned}$$

The tuple $\langle \underline{P}X, \overline{P}X \rangle$ is called a rough set.

Let *P* and *Q* be sets of attributes inducing equivalence relations over $\mathbb{U}$, then the positive region can be defined as:

$$POS_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \underline{P}X \quad (6)$$

The positive region contains all objects of $\mathbb{U}$ that can be classified to classes of $\mathbb{U}$*/Q* using the information in attributes *P*. Based on this definition, dependencies between attributes can be determined. For *P, Q* $\subset \mathbb{A}$, it is said that *Q* depends on *P* in a degree *k* ($0 \leq k \leq 1$), denoted *P* $\Rightarrow_k$ *Q*, if

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|\mathbb{U}|} \quad (7)$$

The reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Attributes are removed so that the reduced set provides the same predictive capability of the decision attribute as the original. A *reduct* R_{min} is defined as a minimal subset *R* of the initial attribute set $\mathbb{C}$ such that for a given set of attributes D, $\gamma_R(\mathbb{D})$ = $\gamma_{\mathbb{C}}(\mathbb{D})$. From the literature, R is a minimal subset if $\gamma_{R-\{a\}}(\mathbb{D}) \neq \gamma_R(\mathbb{D})$ for all $a \in R$.

The main limitation of crisp rough set-based approaches to feature selection is their reliance on nominal data. For processing real-valued data, a discretization step must first be carried out which may result in information loss. This motivates the use of fuzzy-rough sets for feature selection.

2.2. Fuzzy-rough feature selection

Fuzzy-rough sets [6] encapsulate the related but distinct concepts of vagueness (for fuzzy sets) and indiscernibility (for rough sets), both of which occur as a result of uncertainty in knowledge

Definitions for the fuzzy lower and upper approximations can be found in [13], where a $\mathcal{T}$-transitive fuzzy similarity relation is used to approximate a fuzzy concept X:

$$\mu_{\underline{R_P}X}(x) = \inf_{y \in \mathbb{U}} \mathcal{I}(\mu_{R_P}(x, y), \mu_X(y)) \quad (8)$$

$$\mu_{\overline{R_P}X}(x) = \sup_{y \in \mathbb{U}} \mathcal{T}(\mu_{R_P}(x, y), \mu_X(y)) \quad (9)$$

Here, $\mathcal{I}$ is a fuzzy implicator and $\mathcal{T}$ a t-norm. R_P is the fuzzy similarity relation induced by the subset of features P:

$$\mu_{R_P}(x, y) = \mathcal{T}_{a \in P}\{\mu_{R_a}(x, y)\} \quad (10)$$

$\mu_{R_a}(x, y)$ is the degree to which objects x and y are similar for feature a, and may be defined in many ways, for example:

$$\begin{aligned} \mu_{R_a}(x, y) &= 1 - \frac{|a(x) - a(y)|}{|a_{max} - a_{min}|} && (11) \\ \mu_{R_a}(x, y) &= \max(\min(\frac{(a(y) - (a(x) - \sigma_a))}{\sigma_a}, \\ &\qquad \frac{((a(x) + \sigma_a) - a(y))}{\sigma_a}, 0) && (12) \end{aligned}$$

where ${\sigma_a}^2$ is the variance of feature a. As these relations do not necessarily display $\mathcal{T}$-transitivity, the fuzzy transitive closure can be computed for each attribute. The choice of relation is largely determined by the intended application. For feature selection, a relation such as (12) may be appropriate as this permits only small differences between attribute values of differing objects. For classification tasks, a more gradual and inclusive relation such as (11) should be used.

In a similar way to the original crisp rough set approach, the fuzzy positive region can be defined as [9]:

$$\mu_{POS_P(\mathbb{D})}(x) = \sup_{X \in \mathbb{U}/\mathbb{D}} \mu_{\underline{R_P}X}(x) \tag{13}$$

An important issue in data analysis is discovering dependencies between attributes. The fuzzy-rough degree of dependency of $\mathbb{D}$ on the attribute subset P can be defined in the following way:

$$\gamma'_P(\mathbb{D}) = \frac{\sum\limits_{x \in \mathbb{U}} \mu_{POS_P(\mathbb{D})}(x)}{|\mathbb{U}|} \tag{14}$$

A fuzzy-rough reduct R can be defined as a minimal subset of features that preserves the dependency degree of the entire dataset, i.e. $\gamma'_R(\mathbb{D}) = \gamma'_{\mathbb{C}}(\mathbb{D})$. Based on this, a fuzzy-rough greedy hill-climbing algorithm can be constructed that uses equation (14) to gauge subset quality. In [9], it has been shown that the dependency function is monotonic and that fuzzy discernibility matrices may also be used to discover reducts.

3. Unsupervised fuzzy-rough feature selection

This section introduces the new unsupervised subset evaluation measures based on fuzzy-rough set theory, and the corresponding reduction algorithm.

3.1 Dependency measure

The discovery of dependencies between attributes, is in general, an important issue in data analysis. Intuitively, a set of attributes Q depends totally on a set of attributes P, denoted $P \rightarrow Q$, if all attribute values from Q are uniquely determined by values of attributes from P.

The central idea behind the present work is that, as with supervised fuzzy-rough FS [9], the fuzzy dependency measure can also be used to discover the inter-dependency of features. This can be achieved by substituting the decision feature(s) $\mathbb{D}$ of the supervised approach for any given feature or group of features Q such that

$$\gamma'_P(Q) = \frac{\sum\limits_{x \in \mathbb{U}} \mu_{POS_{R_P}(Q)}(x)}{|\mathbb{U}|} \tag{15}$$

where $P \cap Q = \emptyset$ and,

$$\mu_{POS_{R_P}(Q)}(x) = \sup_{z \in \mathbb{U}} \mu_{\underline{R_P}R_Q z}(x) \tag{16}$$

Here, $R_Q z$ indicates the fuzzy tolerance class (or fuzzy equivalence class) for object z. The lower approximation becomes:

$$\mu_{\underline{R_P}R_Q z}(x) = \inf_{y \in \mathbb{U}} \mathcal{I}(\mu_{R_P}(x,y), \mu_{R_Q}(y,z)) \tag{17}$$

3.2. Boundary region measure

Most approaches to crisp rough set FS and all approaches to fuzzy-rough FS use only the lower approximation for the evaluation of feature subsets. The lower approximation contains information regarding the extent of certainty of object membership to a given concept. However, the upper approximation contains information regarding the degree of uncertainty of objects and hence this information can be used to discriminate between subsets. For example, two subsets may result in the same lower approximation but one subset may produce a smaller upper approximation. This subset will be more useful as there is less uncertainty concerning objects within the boundary region (the difference between upper and lower approximations). The fuzzy-rough boundary region for a fuzzy tolerance class $R_Q z$ X may thus be defined:

$$\mu_{BND_P(R_Q z)}(x) = \mu_{\overline{R_P}R_Q z}(x) - \mu_{\underline{R_P}R_Q z}(x) \tag{18}$$

with the upper approximation defined as:

$$\mu_{\overline{R_P}R_Q z}(x) = \sup_{y \in \mathbb{U}} \mathcal{T}(\mu_{R_P}(x,y), \mu_{R_Q}(y,z)) \tag{19}$$

As the search for an optimal subset progresses, the object memberships to the boundary region diminish until a minimum is achieved. From this, the total certainty degree given a feature subset P is defined as:

$$\lambda_P(Q) = 1 - \frac{\sum\limits_{z \in \mathbb{U}} \sum\limits_{x \in \mathbb{U}} \mu_{BND_{R_P}(R_Q z)}(x)}{|\mathbb{U}|^2} \tag{20}$$

It is this measure, λ, that can be used to guide an unsupervised subset selection process.

3.3. Discernibility measure

There are two main branches of research in crisp rough set-based FS: those based on the dependency degree and those based on discernibility matrices and functions. Therefore, it is natural to extend concepts in the latter branch to the fuzzy-rough domain [3].

The fuzzy tolerance relations that represent objects' approximate equality can be used to extend the classical discernibility function. For each combination of features P, a value is obtained indicating how well these attributes maintain the discernibility, relative to another subset of features Q, between all objects.

$$f(P,Q) \quad = \quad \mathcal{T}(\underbrace{c_{ij}(P,Q)}_{1\leq i<j\leq|\mathbb{U}|}) \tag{21}$$

with

$$c_{ij}(P,Q) = \mathcal{I}(\mathcal{T}(\underbrace{\mu_{R_a}(x_i,x_j)}_{a\in P}), \mu_{R_Q}(x_i,x_j)) \tag{22}$$

Alternatively, rather than taking a minimum operation in Eq. (21), one can also consider the average over all object pairs, i.e.,

$$g(P,Q) = \frac{2. \sum\limits_{1\leq i<j\leq|\mathbb{U}|} c_{ij}(P,Q)}{|\mathbb{U}|(|\mathbb{U}|-1)} \tag{23}$$

This measure is less rigid than equation (21), which produces the value 0 as soon as one of the c_{ij} equals 0.

3.4. Finding reductions

For the supervised approach, search is conducted within $\mathcal{P}(\mathbb{C})$, the set of all possible subsets of the conditional feature set. However, for the unsupervised approach search is performed within $\mathcal{P}(\mathbb{C}) \times \mathcal{P}(\mathbb{C})$, as to search for reductions any subset can be compared with any other subset. This is a vastly more complex space in which to search. For the purposes of this paper, a linear backward search is employed that achieves reasonable reductions in a short space of time.

The algorithm (figure 1) starts by considering all of the features contained in the dataset. The removal of each feature is then examined iteratively, and the corresponding measure is calculated. If the measure is unaffected then the feature can be removed. This process continues until all features have been examined. If no interdependency exists, the algorithm will return the full set of features. The complexity for the search in the worst case is $O(n)$, where n is the number of original features.

Reduction is achieved for the three measures by replacing $M(T,\{x\})$ with either $\gamma'_T(\{x\})$, $\lambda_T(\{x\})$ or $g(T,\{x\})$. If a greater reduction in features is required (at the expense of accuracy), line (4) in the algorithm can be replaced by:

if $M(R,\{x\}) < \alpha$

with $\alpha \in (0,1]$.

4. Experimentation

Following feature selection, the datasets are reduced according to the discovered reducts. These reduced datasets are then evaluated using the relevant classifier learning method (as described below) and evaluated with 10-fold cross validation.

```
UFRQUICKREDUCT($F$)
$F$, the set of all features.

(1)  $R \leftarrow \mathbb{C}$
(2)  foreach $x \in \mathbb{C}$
(3)      $R \leftarrow R - \{x\}$
(4)      if $M(R,\{x\}) < 1$
(5)          $R \leftarrow R \cup \{x\}$
(6)  return $R$
```

Figure 1. The UFRQUICKREDUCT Algorithm

Two learning mechanisms were employed to create classifiers for the purpose of evaluating the resulting subsets from the feature selection phase: JRip [2] and J48 [16]. JRip learns propositional rules by repeatedly growing rules and pruning them. During the growth phase, features are added greedily to fit training samples. Once the ruleset is generated, a further optimisation is performed where rules are evaluated and deleted, based on their performance on randomised data. J48 creates decision trees by choosing the most informative features via an entropy measure, and recursively partitions the data into subtables based on their values. Each node in the tree represents a feature with branches from a node representing the alternative values this feature can take according to the current subtable. Partitioning stops when all data items in the subtable have the same classification.

The feature selection methods employed are: correlation-based (CFS) [7], fuzzy-rough lower approximation-based (FRFS), boundary region-based (B-FRFS), discernibility-based (D-FRFS), unsupervised fuzzy-rough lower approximation-based (UFRFS), unsupervised boundary region-based (B-UFRFS) and unsupervised discernibility-based (D-UFRFS). The classification accuracies for the unreduced data are also included for comparison. All of the data used in this experimental investigation is labelled. However, before applying the unsupervised methods, the decision feature is removed from the data, and the approaches operate on the unlabelled data only. When learning classifiers, or applying supervised FRFS, the complete dataset is used.

The results presented in Table 1 show the subset sizes discovered by the methods. As feature selection takes place for each fold in the cross-validation, the subset sizes in the table are averages. It can be seen that the proposed methods manage reduction in all cases and return substantial levels of dimensionality reduction for some datasets. These results compare well with the supervised approach and show that the unsupervised approaches may even find smaller subsets

Table 1. Subset sizes for UFRFS

Dataset	Features	Objects	CFS	FRFS	B-FRFS	D-FRFS	UFRFS	B-UFRFS	D-UFRFS
Cleveland	13	297	6.7	7.7	7.7	7.7	10.5	10.5	10.5
Glass	9	214	6.3	9	8.2	8.2	7.1	7.1	7.1
Heart	13	270	7.4	7.1	7.1	7.1	10.2	10.2	10.2
Ionosphere	34	230	15	5	5	5	6.2	6	6.2
Olitos	25	120	10.8	7.1	7.1	6.9	9	9	9
Water 2	38	390	9.1	6	6	6	7	7.4	7
Water 3	38	390	10.6	6	6	5.9	7.1	7.4	7.1
Web	2556	149	54.4	18.4	17.4	16.0	18.4	18.8	18.4
Wine	13	178	10.7	5	4.9	4.8	6	6.2	6

in some cases.

The resulting classification accuracies for the classifiers can be seen in Table 2 and Table 3. These demonstrate that the unsupervised methods retain useful features, without considering the decision feature. This is borne out by comparison to the classification accuracy of the unreduced data, showing that the greatest decrease amongst all of the reduced data is only in the order of 10% overall. There are also cases where the use of unsupervised-reduced data outperforms the unreduced data and that of the supervised-reduced data.

In table 2, there are only five cases where the drop in accuracy is statistically significant. For the `Web` data, the performance of UFRFS and D-UFRFS were statistically worse and for the `Wine` data, the drop in accuracy for the three unsupervised methods was significant. In table 3, there is only one dataset (`Wine`) for which the unsupervised methods perform statistically worse than the unreduced approach. This demonstrates the power of the unsupervised methods as they perform drastic dimensionality reduction that generally maintains the classification accuracy whilst ignoring the class information. This implies that the quality of reduction should be high for datasets with no class labels using these techniques.

5. Conclusion

This paper has presented novel techniques for unsupervised feature selection, based on the fuzzy-rough dependency measure. These approaches are data-driven, and no user-defined thresholds or domain-related information is required, although a choice must be made regarding fuzzy similarity relations and connectives. Note that these choices must also be made for the existing supervised FS approaches that employ the same underlying mathematical theory. The results show that the approach can reduce dataset dimensionality considerably whilst retaining useful features.

At present the unsupervised search algorithm utilises a simple but nevertheless effective backwards elimination method for search. The problem with such search techniques is that they often return a result which is sub-optimal. The investigation of other search techniques such as ant colony optimisation [8] and particle swarm optimisation [14], may help in alleviating this problem and thus further improving the efficiency of the approaches. Also, a more complete comparison of UFRFS and other unsupervised FS techniques for clustering performance, would form the basis for a series of topics for future investigation.

As mentioned previously the fuzzy similarity relations and connectives must be chosen for UFRFS. As only one choice of fuzzy connective (Łukasiewicz), and also a single fuzzy similarity measure(as defined in 11) are explored in this paper, the evaluation of other options in this regard would form the basis for a further more comprehensive investigation.

A further interesting topic is the use of a method which would combine both the unsupervised, and supervised measures. The supervised measure determines relevance and the unsupervised measure determines redundancy, hence a method that combines these should be particularly powerful for subset evaluation.

References

[1] C.L. Blake, C.J. Merz, *UCI Repository of Machine Learning Databases.* Irvine, University of California, 1998. http://archive.ics.uci.edu/ml/

[2] W.W. Cohen, "Fast effective rule induction," In *Proceedings of the 12th International Conference on Machine Learning*, pp. 115–123, 1995.

[3] C. Cornelis, G. Hurtado Martín, R. Jensen, D. Ślęzak, "Feature Selection with Fuzzy Decision Reducts," *3rd Int. Conf. on Rough Sets and Knowledge Technology (RSKT'08)*, pp. 284–291, 2008.

[4] M. Dash and H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis*, vol. 1, no. 3, 1997.

Table 2. Classification accuracies: JRip (%)

Dataset	Unred	CFS	FRFS	B-FRFS	D-FRFS	UFRFS	B-UFRFS	D-UFRFS
Cleveland	52.18	56.93	54.52	54.52	54.52	54.90	53.57	54.90
Glass	71.39	68.23	71.39	66.90	66.90	64.94	64.94	64.94
Heart	77.41	78.89	77.04	77.04	77.04	80.00	80.74	80.00
Ionosphere	70.83	71.67	65.83	69.17	63.33	60.00	67.50	60.00
Olitos	86.52	90.87	86.52	86.52	85.22	86.09	84.78	86.09
Water 2	82.82	82.31	84.10	84.10	81.28	84.87	82.82	84.87
Water 3	81.79	82.56	78.97	80.26	81.28	80.00	78.97	80.00
Web	58.38	55.05	50.38	48.33	47.76	47.05	47.76	47.05
Wine	95.00	92.12	88.27	91.57	90.46	74.67	79.90	74.67

Table 3. Classification accuracies: J48 (%)

Dataset	Unred	CFS	FRFS	B-FRFS	D-FRFS	UFRFS	B-UFRFS	D-UFRFS
Cleveland	51.87	56.92	49.84	49.84	49.84	52.91	50.21	52.91
Glass	67.29	69.98	67.29	65.87	65.87	65.91	65.91	65.91
Heart	76.67	80.74	77.04	77.04	77.04	78.89	79.26	78.89
Ionosphere	67.50	57.50	62.50	61.67	70.83	59.17	64.17	59.17
Olitos	87.83	88.70	86.96	86.96	86.52	85.22	83.91	85.22
Water 2	83.08	84.10	84.36	84.36	83.59	83.59	82.05	83.59
Water 3	83.08	81.54	79.49	80.26	80.77	81.54	80.51	81.54
Web	51.57	59.71	47.05	48.38	42.29	45.05	46.33	45.05
Wine	94.41	94.41	94.97	96.08	94.41	79.74	81.99	79.74

[5] M. Dash and H. Liu, "Unsupervised Feature Selection," Proceedings of the Pacific and Asia Conference on Knowledge Discovery and Data Mining, pp. 110–121, 2000.

[6] D. Dubois, H. Prade, "Rough fuzzy sets and fuzzy rough sets," *International Journal of General Systems*, vol. 17, 91–209, 1990.

[7] M.A. Hall, "Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning," Proceedings of the 17th International Conference on Machine Learning, pp. 359–366, 2000.

[8] R. Jensen, Q. Shen, *Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches*, IEEE Press/Wiley & Sons, 2008.

[9] R. Jensen, Q. Shen, "New approaches to fuzzy-rough feature selection," *IEEE Transactions on Fuzzy Systems*, in press, 2009.

[10] P. Mitra, C.A. Murthy, and S.K. Pal, "Unsupervised Feature Selection Using Feature Similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 1–13, 2002.

[11] S.K. Pal, R.K. De, and J. Basak, "Unsupervised Feature Evaluation: A Neuro-Fuzzy approach," *IEEE Transactions in Neural Networks*, vol. 11, pp. 366–376, 2000.

[12] Z. Pawlak, "Rough sets," *International Journal of Computing and Information Sciences*, vol. 11, pp. 341–356, 1982.

[13] A.M. Radzikowska and E.E. Kerre, "A comparative study of fuzzy rough sets," *Fuzzy Sets and Systems*, vol. 126, no. 2, pp. 137–155, 2002.

[14] X. Wang, J. Yang, X. Teng, W. Xia and R. Jensen, "Feature Selection based on Rough Sets and Particle Swarm Optimization," *Pattern Recognition Letters*, vol. 28, no. 4, pp. 459–471, 2007.

[15] I.H. Witten and E. Frank, "Generating Accurate Rule Sets Without Global Optimization," *Proceedings of the 15th International Conference on Machine Learning*, Morgan Kaufmann Publishers, San Francisco, 1998.

[16] I.H. Witten and E. Frank, *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, 2000.

[17] W. Ziarko, "Variable Precision Rough Set Model," *Journal of Computer and System Sciences*, vol. 46, no. 1, pp. 39–59, 1993.