International Journal of Hybrid Information Technology
Vol.9, No.1 (2016), pp. 119-128
http://dx.doi.org/10.14257/ijhit.2016.9.1.11

An Effective Feature Selection Approach Using the Hybrid Filter Wrapper

Haitao Wang[1] and Shufen Liu[2]

[1]*School of Computer Science and Technology*
Henan Polytechnic University
GaoXin District, JiaoZuo City, HeNan Province, China
E-mail: Jz_wht@126.com
[2]*School of Computer Science and Technology*
JiLin University
QianJin Street, ChangChun City, JinLin Province, China
Jz_wht@126.com

Abstract

Feature selection is an important data preprocessing technique and has been widely studied in data mining, machine learning and granular computing. In this paper, we introduced an effective feature selection method using the hybrid approaches, that is, use the mutual information to select the candidate feature set, then, obtain the super-reduct space from the candidate feature set by a wrapper search approach, finally, the wrapper method determined to select the proper features. The experimental results show that our approach owned the obvious merits in the aspect of classification accuracy ratio and number features selected by extensive comparing with other methods.

Keywords*: feature selection, super-reduct, mutual information, classification*

.

1. Introduction

In today's world, huge amount of data is produced electronically by various sources such as social group network, stock exchanges, web pages, emails, reports in industries and newsgroup. A number of intelligent information systems such as product categorization, search personalization, sentiment mining, document filtering for digital libraries, authorship detection, product review classification, email filtering, and biographic generation are built over text documents on the web. Document categorization is the preliminary step in building such intelligent system. The task of document categorization involves understanding of both the content of documents and the role of the categories. Due to the tremendous increase in electronic data, machine learning has become the principle approach to document categorization. Machine learning algorithm build a model form the training corpus by observing the characteristics of the documents under each category. Document representation is the primary course for any machine learning algorithm.

Feature selection is a data preprocessing course that decreases the dimensionality by removing such redundant and irrelevant data [1]. It offers a better understanding of the data by giving only the important feature. When feature selection is processed well, the computational and memory demands of the inducer and the predictor are reduced and the accuracy of the classifier gets improved. A number of feature selection algorithms have been proposed and their efficiencies have been studied for text categorization.

Feature selection plays a key role in classification task. Irrelevant and redundant features in input features not only complicate the problem but also degrade solution

ISSN: 1738-9968 IJHIT

accuracy. The aim is to select the essential features that allow us to discern between patterns belonging to different classes. In all feature selection method, we can divide feature selection methods into two categories: filter method and wrapper method. Filter methods select a subset of features as preprocessing step which is independent from the induction algorithm. The best feature subset is selected by evaluating some predefined criterion without involving any learning algorithm. Therefore, this concept usually considers a faster speed importantly. Moreover, the filter method [2] is computational less expensive and more general. Wrappers utilize the performance of the classifier accuracy for a particular classifier at cost of high computational complexity and less generalization of the selected features on the other classifier. However, the wrapper method generally outperforms the filter method in the aspect of the accuracy of the learning machine.

To these issues above, we proposed the super-reduct wrapper, a hybrid filter/wrapper method to obtain the feature set, our contribution is mainly the feature selection approach by using the hybrid model to obtain relative high classification accuracy and less number of features.

The rest of this paper is organized as follows. Section 2 introduces the related work in the aspect of feature selection, Section 3 presents the effective feature selection course and detail algorithm that utilized the filter wrapper and mutual information. Experimental result and analyses are described in Section 4. Finally, conclusion are given in Section 5.

2. Related Work

As mentioned previously, a large number of filter-based feature selection algorithms have been presented in the past few decades for mining the optimal features subset from the high dimensional feature spaces [3]. The feature ranking based selection methods have been presented to calculate the features scoring based on constructing the different scoring function. The variance score method is to select the features with maximum variances by calculating the variance of each feature to reflect its representative power. The Lplacian Score method is another popular unsupervised feature filter selection under the assumption that two close samples should have similar feature values and a good feature margin values for samples in different classes. Using normalized mutual information to measure the dependence between a pair of features, Q. Hu, D. Yu, proposed a heterogeneous feature selection method based on Neighborhood rough set. A number of supervised learning algorithm have been used to implement the filter methods, which included Relief family and Fuzzy-Margin-based relief. Batti investigated features and to select the top ranked feature used as input data for neural network classifier. These algorithms based on score function are widely used in data mining and pattern recognition [5-8]. However, the defect of these methods lies on ignoring the redundant features and bring a bad influence on the performance of the following classifiers. To overcome the above problem, the optimal feature subsets have been affected considering the redundancy among the selected features by most researchers. Mutual Information (MI) is a measure of the amount of information between tow random variables, which is symmetric and non-negative, and is zero if and only if the variables are independent. Then, the methods based on MI have been popular lately [9]. Yu and Liu introduced a novel framework that decoupled relevance analysis and redundancy analysis. They proposed a correlation-based subset selection method named FCBF for relevance and redundancy analysis, and then removed redundant features by approximate to Markov Blanket technique. The MIFS algorithm was proposed to calculate the mutual information both with respect to class variables and the already selected features for each feature and selected

those feature hat have maximum mutual information with class labels but less redundant among the selected feature. However, MIFS algorithm ignored feature synergy and it variants may cause a big bias when feature are combined to cooperate together. To avoid drawbacks, W.-Z. Wu, Y. Leung, proposed a novel feature selection method for optimal scale selection for multi-scale decision tables, aiming at greatly relieving the computation overhead and a set of individually discriminating and weakly dependent features can be selected. Based on information gain and MI, FESLP was proposed by Ye Xu to address the link prediction problem, whose superior advantage is that those feature with the greatest discriminative power are selected and simultaneously the correlations among features such that redundancy in the learned feature space are minimized as small as possible.

The classical rough set is popular for finding a subset (named a reduct) of the original attributes that are most informative. The feature selection algorithm mostly use the properties of rough set such as core attributes, projection operations, and dependency of attributes to support the finding of reduct. After that, one attribute selected at each iteration is added until the reduct contains the same quality of classification as the original. One problem with the classical rough set is that it may fail to find an optimal reduct with noisy dataset, some reducts with more feature can perform good while some with fewer features is possible to perform bad. So, it's crucial to select the reduct that is suitable for a particular learning machine and with highest classification accuracy. Variable precision rough set model (VPRSM) is an extension of the classical rough, which uses the concept of majority inclusion which is a proportion of an object in condition class belonging to a decision class for a given classification.

Filter approach is also feature selection method which is widely utilized and two key reasons are spending less expensive computation for the data with a large number of features and its generality. Concerning it's classification accuracy, the performance of the learning algorithms, however, degraded in some situations. This is because the filter model separates feature selection form the classifier learning and selects the feature subsets that are independent from learning algorithm. Therefore, efficiency and effectiveness of these methods depend on the subset evaluation used to measure the goodness of a feature subset in determining an optimal one. It relies on various measures of general characteristics of the training data. As the described above, the filter approach is efficient for high dimensional data due to its linear time complexity in terms of dimensionality. However, it separates feature selection from the classifier learning and selects feature subsets that are independent from any learning algorithm because of none of learning algorithm involved [10-12], therefore, many method based filter approach may produce insufficient predictive accuracy when applying to some learning algorithm, moreover, it's unnecessary to select too many feature because of feature ranking in course of feature selection.

The wrapper model utilizes the classification accuracy of a predetermined learning algorithm to determine the goodness of selected subsets. It searches for features that better suited the learning algorithm, aiming to improve the performance of the learning algorithm, but it's more computationally expensive than the filter models. The feature selection algorithm exists around the learning algorithm. The learning algorithm runs on the datasets by the subsets of the features, and the subset of feature with the highest classification accuracy is chosen. Therefore, the wrapper approach generally outperforms the filter approach in the aspect of the final predictive accuracy of learning machine. However, this model aims to only one classifier and owns more expensive in computational cost. To overcome the limitations of these aspects, some strategies proposed to conduct feature selection, which were independent form the machine learning and used the

classification accuracy of learning algorithm to only determine the goodness of the subsets selected. So, how to decrease the computational complexity and cost and obtain a few feature with high predictive accuracy at the same time, these aspects are research goals which we always dedicated.

3. Course of Feature Selection

Our goal is to design the efficient algorithm in selection the best set of features and reduce the search space of the wrapper mode to decrease computational cost. Based on these criteria, we present a novel feature selection algorithm contained three steps: firstly, find a candidate feature set using an incremental selection method, then, obtain the suited super-reduct by special algorithm we proposed to reduce the search space from a candidate feature set and avoid the local maximal problem, finally, using the wrapper model with the sequential backward elimination scheme to search the proper reduct from the superreduct.

3.1 Find a Candidate Feature Set

The main objective of the feature selection is to obtain a feature subset consisting of low dimensionality, sufficient information preserving and improvement of classification accuracy by removing impacts on the irrelevant and redundant features. In our research, we use conditional mutual informational criterion for ranking the feature and subset of features is selected from the top of a ranking list, which approximates the set of the relevant features. This approach is efficient for high-dimensional data with its linear time complexity in terms of dimensionality N. However, the selected features may often contain many redundant features because the redundant features because the redundant features are likely to have high relevance with respect to the decision variable. Hence, we roughly refine the redundant and irrelevant features of the candidate feature set with the superreduct in order to decrease the time complexity of the searching for the best feature subset on the wrapper approach. To select the candidate feature set, we compute the classification accuracy for a ranking feature and select the subset of features with highest classification accuracy on the test set, which reason is that we desire a powerful subset of the feature and it’s likely to be efficient.

3.2 Select the Best Superreduct

As discussed above, it’s impossible to find all of the reduct for dataset. In this section, we proposed an algorithm 1 for approximating the multiple superreducts on the candidate feature set. Algorithm 1 is to find the set of the approximate superreduct on the candidate feature set, which presented as following:

Produce1:

Input: Decision information system IS = (U, C∪D), B ⊆ C.
Output: Superreduct R which is cater to the quality of classification γ(C,D, μ).
Step1: set R = {}
Step2: calculate the quality of classification γ(C,D, μ) by **procedure 2**.
Step3: **while** γ(C,D, μ) + ∈ < γ(C,D, μ) do
  for all a ∈ C - R do
    select a that yields the largest γ(R∪{a},D, μ) + θ(R∪{a},D, μ)
  end for
  if γ(R∪{a},D, μ) > γ(R,D, μ) then
    R = R ∪{a}
  else
    R = {Φ}, go to step4.

Algorithm 1: Multiple superreduct algorithm

Input: the candidate feature set (*CF*) and decision information system *IS*=(U,C∪D), *CF*⊆ C, and B ⊆ C
Output: set of superredcts RS
Step1: set Rs = {}
Step2: **for all** a ∈ *CF* do
  set B = {}
  B = B∪{a}
  for all e ∈ *CF* – B do
    append e to B
  end for
  calculate the approximate superreduct for set B by **produce1** and append that superreduct to Rs
end for
Step3: removing the redundant superreduct form Rs
Step4: Return Rs

Produce2:

Input: Decision information system IS = (U, C∪D), B ⊆ C.
Output: Maximum quality of classification (γ_{max}) and its related μ.
Step1: set $\mu = 0.3 + \varepsilon_1$
Step2: calculate the quality of classification γ(B,D, μ) and set $\gamma_{max} = \gamma(B,D, \mu)$, $\sigma_{max} = \sigma(B,D, \mu)$.
Step3: **while** μ ⩽ 1 do
  $\mu += \varepsilon_2$
  calculate the γ = γ(B,D, μ), σ = σ(B,D, μ)
  if $\gamma > \gamma_{max}$ **then**
    $\gamma_{max} = \gamma$, $\sigma_{max} = \sigma$
  else
    if $\gamma = \gamma_{max}$ then
      if $\sigma < \sigma_{max}$ then
        $\gamma_{max} = \gamma$, $\sigma_{max} = \sigma$
      end if
    end if
  end if
end while
Step4: **return** γ_{max}

For the candidate feature set, n features, the number of the superreduct is less than or equal to n. Therefore, the time needed to compute the multiple superreduct is more than one reduct. However, the main advantages for finding the multiple super-reducts instead of one reduct are that (1) it has more generality than one reduct, (2) it increases the opportunity to investigate the best superreduct which is more suitable for learning algorithm. (3) It can breakdown the limitation of one super-reduct that encounters the local maxima, and (4) it can find the dependencies between the groups of feature.

Similar to other methods, a train set is used for train predetermined MLPs and the performance of the selected features is tested on the test set. In this aspect, the super-reduct with highest classification accuracy on the test set is chosen as the best-super-reduct. In addition, the best-superreduct will be used as feature space selecting as subset of features by the wrapper approach. In the next section, we show the average accuracy

value between the results obtained through the three approaches based on the original feature set, the CMI, the Superreduct-Wrapper. In some cases, these accuracy differences are not statistically significant. However, it is worth to emphasize that the subset selected obtained by superreduct-Wrapper has fewer number of feature.

3.3 Selecting the Best Feature Subset on Wrapper Approach

As we known, the learning algorithms are used to control the selection of feature subsets, the wrapper model tends to give superior performance as feature subsets are found better suited the predetermined learning algorithms, Consequently, it's also more computationally expensive that filter mode. In addition, selecting a subset of feature on the wrapper model, the time complexity is quadratic in terms of data dimensionality for the sequential search. Therefore, it's not proper for the dataset better suited the learning algorithm which deals with a large number of candidate feature. Moreover, at each pass, it searches for every feature in the candidate features to find the feature that reduces as many as error. So, if the number of features in the candidate feature is k, the determination is done by the different k configurations. Thus, this method has high computational cost when applying to the datasets with a large number of the features.

In this aspect of work, we proposed the reducing of the search space of the candidate feature set to the best super-reduct which reduces the computational cost of the wrapper search. Our approach utilized the sequential backward elimination technique to search for every possible subset of features through the best super-reduct space. The features ranked according to the average accuracy of classifier, and then feature will be removed one by one from the best only if such exclusion improves or does not change the classifier accuracy. In addition, in some cases, our method is unnecessary to search for every possible feature of the best superreduct as the complete search does. The decremental selection procedure for selecting a proper reduct on the wrapper model can be shown in Algorithm 2.

There are two phases in the algorithm, named superreduct-wrapper, in the first stage, the features are ranked determining from the average accuracy of the classifier, namely step 4, in the second stage, we deal with the list of the ordered features once, each feature in the list determines the first till the last ranked feature, namely step 6- 12. In this stage, each feature in the list considers the average accuracy of the classifier only if the feature excluded. If any feature is found to lead to the most improved average accuracy and the relative accuracy is more than δ_1 (step 7), the feature then will be removed. Otherwise, every possible feature is considered and the feature that leads to the largest average accuracy will be chosen and removed (step 8). The one that leads to the improvement or the unchanging of the average accuracy (step 9), or the degrading of the relative accuracy not worst than δ_2 (step 10) will be removed. This decremental selection procedure is repeated until the termination condition is met.

Usually, the sequential backward elimination is more computationally expensive than the incremental sequential forward. However, it could yield a better result when considering the local maximal. In addition, the sequential forward search adding one feature at each pass doesn't take the interaction between the groups of the features into account. In many classification problem, the relevance of the features may be grouped by several features arising acting simultaneously but not the individual feature alone.

Superreduct_Wrapper algorithm

Input: D_{train}, D_{test}, Superreduct
Output: Obtain an ideal reduct B
Step1: Classifier(D_{train}, Superreduct)
Step2: Acc_{sr} = (D_{test}, Superreduct)
Step3: set B = {}
Step4: **for all** $f_i \in$ *Superreduct* do
Classifier(D_{train}, f_i)
Append f_i to B
end for
Step5: put the feature in B in an ascending order according to Score value
Step6: **while** $|B| > 0$ **do**
for all $f_i \in B$ according to order **do**
Classifier(D_{train}, B-{ f_i })
Acc_{fi} = Classifier(D_{test}, B-{ f_i })
if (Acc_{fi} - Acc_{sr}) / $Acc_{sr} > \delta_1$ **then**
$B = B$-{ f_i }, $Acc_{sr} = Acc_{fi}$
go to step 6
end if
select f_i with the maximum Acc_{fi}
end for
if ($Acc_{fi} \geq Acc_{sr}$) **then**
$B = B$-{ f_i }, $Acc_{sr} = Acc_{fi}$
go to step 6
end if
if (Acc_{sr} - Acc_{fi}) / $Acc_{sr} \leqslant \delta_2$ **then**
$B = B$-{ f_i }, $Acc_{sr} = Acc_{fi}$
go to step 6
end if
Step7: go to step9
Step8: **end while**
Step9: return the best of the selected feature subset in B

4. Experimental Results and Analyze

This section illustrates the evaluation of our method in terms of the classification accuracy and the number of features in order to detect how good the method our presented is in the situation of large and middle-sized features. In addition, the performance of the CMI algorithm was compared with the results of MIFS, MIFS-U, and mRMR, to illustrate the efficiency and effectiveness of our method. Four datasets from the UC-Irvine repository shown in Table 1 were used to assess the performance of the algorithm our offered.

In all case as the following described, the MI was estimated by using the equidistance partitioning for continuous features. The redundancy parameter for MIFS and MIFS-U was varied in the range between 0.0 and 1.0 with a step size of 0.1. The results obtained with best value are used for comparing with CMI. Moreover, in our research a MPL with single hidden layer was trained by using the back-propagation algorithm, in order to evaluate the merits of the feature subsets. The classification accuracies presented here are average of ten trails with random initializations. All datasets were divided into two main groups with three disjointed set because the testing results of some datasets may be low. First group, data sets were partitioned as training (50%), validation (25%), and testing (25%), the second group consists of training (70%), validation (10%), and testing (20%). The minimum error rate of the correct classifications on the validation set was used as a stopping criterion. Selecting the optimal number of the hidden units was considered by

running the BP algorithm in the range between 1 and 3, and the MLP architecture with best validation results was chosen.

Table.1 Details of Datasets

No	DataSets	Instances	#Fea.	Classes
1	Sonar	208	60	2
2	Musk1	476	168	2
3	Arrhythmia	452	279	2
4	Vehicle	846	18	2
5	Spectf	267	44	2

As already mentioned, the proposed method reduced the search space before the wrapper search used to select the best subset of the feature that suited the learning algorithm. The first stage is to select the candidate feature and the subset of the features selected by CMI algorithm with highest accuracy is chosen. The second is to reduce the computational effort of the wrapper search where the candidate feature set is reduced to be many super-reduct and select the super-reduct with the highest classification accuracy as the best-superreduct. Last stage, the best superreduct is used for the search space of the wrapper method in order to further search for the subset of the features that most suited the learning algorithm.

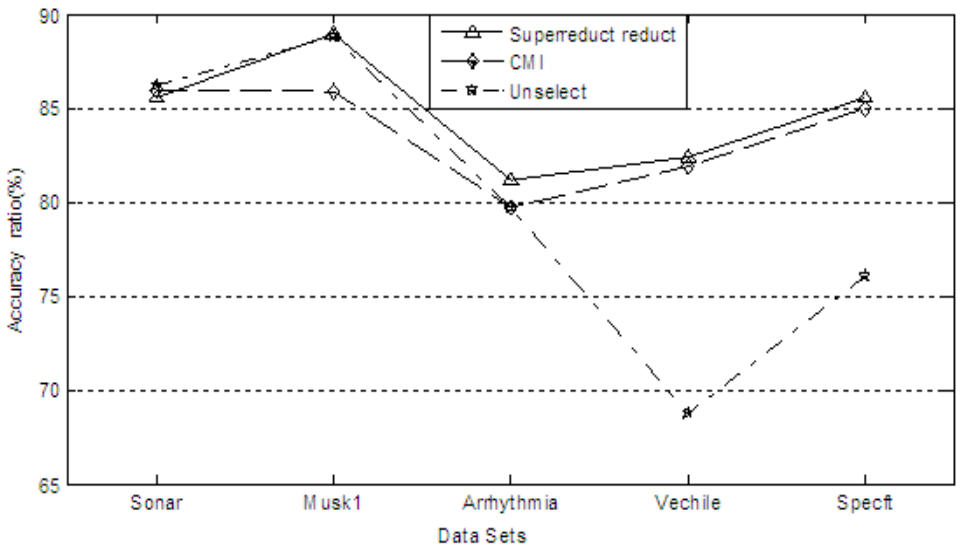

Figure 1. Accuracy Ratio on Various Data Sets

The results in Figure 1 and Figure 2 show that the classification accuracies and the number of the selected features obtained from super-reduct wrapper, CMI method and unselect. As we can see that the accuracy results obtained from the super-reduct wrapper method is generally better than those from the CMI on datasets, at the same, is also better than with the unselect. Moreover, the classification accuracies of super-reduct wrapper on average have more numbers of the maximum values than the CMI and the unselect. However, in all case, the numbers of features selected by the super-reduct wrapper are lower than other methods.

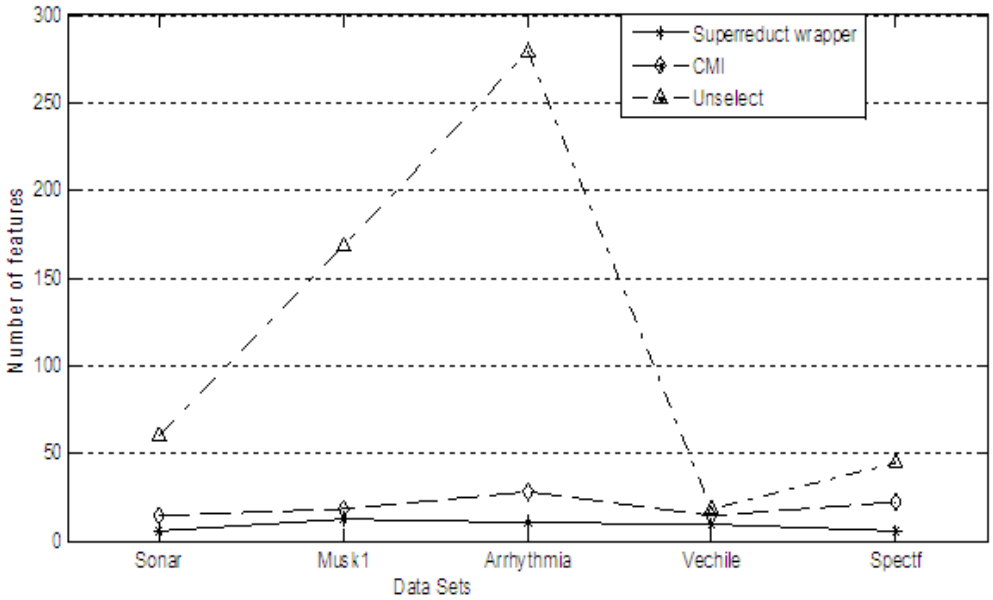

Figure 2. Number Features on Various Data Sets

5. Conclusion

Traditional feature selection methods with the wrapper approach often select on a set of all features or the candidate features which are all large feature, however, this inclusion of irrelevant, redundant and noisy features in the search space can result in poor predicative performance and increase computation cost, moreover, in some case wrapper search may select a subset of the feature that is local maxima and may fall into a trap. To these issues above, we proposed the super-reduct wrapper, a hybrid filter/wrapper method to find a proper reduct suited the learning algorithm. A superreduct was used as the search space for finding a proper reduct where either the irrelevant or redundant features were eliminated according to determining the classifier, at the same time, experimental results shown that superreduct wrapper chose a small subset of features from a super-reduct and provided good performance in the aspect of predicative accuracy comparing to the other methods. Of course, how to reduce the complexity and increase the accuracy degree of classification of our method is our dedicated research direction in the future work.

Acknowledgement

The authors are grate to the editor and anonymous reviewers for their valuable comments on this paper, and the work of this paper is supported by the National Nature Science Foundation of China (No.61300216), the National Innovative Approach Special Project (No.2012IM010200), the Science and Technology Research Projects of HeNan Province (No. 132102210123), the Open Laboratory Project of Mine Information Key discipline of Henan University.

References

[1] Y. Du, Q. Hu, P. Zhu and P. Ma, "Rule learning for classification based on neighborhood covering reduction", Inform. Sci., vol. 181, **(2011)**, pp. 5457-5467.
[2] M. E. ElAlami, "A filter model for feature subset selection based on genetic algorithm", Knowledge-Based Systems, vol. 22, no. 5, **(2009)**, pp. 356-362.
[3] Q. Hu, J. Liu and D. Yu, "Mixed feature selection based on granulation and approximation", Knowledge -Based Syst., vol. 21, **(2008)**, pp. 294-304.
[4] Q. Hu, D. Yu, J. Liu and C. Wu, "Neighborhood rough set based heterogeneous feature selection", Inform. Sci., vol. 178, **(2008)**, pp. 3577-3594.

[5] Q. Hu, W. Pan, L. Zhang, D. Zhang, Y. Song and D. Yu, "Feature selection for monotonic classification", IEEE Trans. Fuzzy Syst., vol. 20, **(2012)**, pp. 69-81.
[6] X. Jia, W. Liao, Z. Zhen and L. Shang, "Minimum cost attribute reduction in decision theoretic rough set models", Inform. Sci., vol. 219, **(2013)**, pp. 151–167.
[7] J. Liang, F. Wang, C. Dang and Y. Qian, "An efficient rough feature selection algorithm with a multi-granulation view", Inter. J. Approx. Reason, vol. 53, **(2012)**, pp. 912-926.
[8] G. Lin, Y. Qian and J. Li, "NMGRS: neighborhood-based multi granulation rough sets", Int. J. Approx. Reason, vol. 53, **(2012)**, pp. 1080-1093.
[9] Y. Lin, J. Li and S. Wu, "Rough set theory in interval and set-valued information systems", Control Decis., vol. 26, **(2011)**, pp. 1611-1615. in Chinese.
[10] W.-Z. Wu and Y. Leung, "Optimal scale selection for multi-scale decision Tables", Int. J. Approx. Reason, vol. 54, **(2013)**, pp. 1107-1129.
[11] S. S. Kannan and N. Ramaraj, "A novel hybrid feature selection via symmetrical uncertainty ranking based local search algorithm", Knowledge-Based Systems, vol. 23, no. 6, **(2010)**, pp. 580-585.
[12] S. Mohammed and O. Mohammed, "Classifying Unsolicited Bulk Email (UBE) using Python Machine Learning Techniques", International Journal of Hybrid Information Technology, vol. 6, no. 1, **(2013)**, pp. 43-48.

Authors

Haitao Wang, He was born in 1974.10, Ph.D. vice professor, major in computer system architecture, his research interests includes cloud computing, parallel computing, data mining, high performance computing.

Shfeng Liu, She was born in 1950, professor, Ph.D. supervisor, research on computer cooperative work, clouding computing, simulation modeling, preside over national 863 key project, supportive project, major special projects and so on.

