An Efficient Fuzzy Rough Approach for Feature Selection

Feifei Xu[1], Weiguo Pan[1], Lai Wei[2], and Haizhou Du[1]

[1] Shanghai University of Electric Power
Shanghai 200090, China
xufeifei1983@hotmail.com, panweiguo@shiep.edu.cn, du_hz@126.com
[2] College of Information Engineering, Shanghai Maritime University
Shanghai 201306, China
weilai@shmtu.edu.cn

Abstract. Rough set theory is a powerful tool for feature selection. To avoid the information loss by discretization in rough sets, fuzzy rough sets are used to deal with the continuous values. However, the cost of computation of the approach is too high to be worked out as the number of selected features increases. In this paper, a new computational method is proposed to approximate the conditional mutual information between the selected features and the decision feature, and thus improve the efficiency and decrease the complexity of the classical fuzzy rough approach based on mutual information. Extensive experiments are conducted on the large-sized coal-fired power units dataset with steady state, and the experimental results confirm the efficiency and effectiveness of the proposed algorithm.

Keywords: Fuzzy rough sets, Feature selection, Mutual information, Large-sized coal-fired power units.

1 Introduction

Feature selection has been a fertile field of research and development since 1970's and proven to be effective in removing irrelevant and redundant features, increasing efficiency in learning tasks, improving learning performance like predictive accuracy, and enhancing comprehensibility of learned results[1]. Fuzzy-rough feature selection (FRFS) can be applied to data with continuous or nominal attributes, and as such can be applied to regression as well as classification datasets[2]. However, due to the computation complexity of feature selection algorithms based on fuzzy rough approach, research on fuzzy rough feature selection is challenging[2–4]. A prominent method for fuzzy rough feature selection is based on Pawlak' algebra[5], proposed by R. Jensen and Q. Shen[2, 4]. Additionally, from the view of information theory, a mutual information-based algorithm for fuzzy rough feature selection (MIFRFS)was proposed[6].

However, MIFRFS algorithm is computationally very costly as the number of selected features is huge. The problem is primarily striking in the practical application. In attempt to attack this problem, in this paper, an efficient

J.T. Yao et al. (Eds): RSKT 2011, LNCS 6954, pp. 95–100, 2011.

feature selection method based on mutual information is proposed by using a novel criteria. It employs fuzzy rough approach to compute the relevance and independence of the features, instead of calculating the conditional mutual information between the selected features and the decision feature(Cartesian product of the fuzzy classes among all selected features), which immensely decreases the computational complexity while maintaining high predictive accuracy. The performance of the proposed approach is compared with that of classical fuzzy rough approach using the predictive accuracy of 1-nearest neighbor (1-NN) rule on large-sized coal-fired power units dataset with steady state.

2 Mutual Information-Based Algorithm for Fuzzy Rough Feature Selection (MIFRFS)

Fuzzy rough sets combine the advantages from both rough sets and fuzzy sets. The information-theoretic expression of knowledge in fuzzy rough sets is briefly introduced as follows.

Definition 1. *Suppose* $U = \{x_1, x_2, ..., x_N\}$*, fuzzy attribute set* $\tilde{A}$ *is composed of a group of fuzzy attributes* $\{\tilde{A^1}, \tilde{A^2}, ..., \tilde{A^M}, \tilde{A^{M+1}}\}$*.* $D = \{\tilde{A^{M+1}}\}$ *is a fuzzy decision attribute. Others are fuzzy condition attributes* $C = \{\tilde{A^1}, \tilde{A^2}, ..., \tilde{A^M}\}$*. Each fuzzy attribute can partition the* U *into* p_j *fuzzy equivalence classes, namely,* $F(\tilde{A^j}) = \{\tilde{F_1^j}, \tilde{F_2^j}, ..., \tilde{F_{p_j}^j}\}(j = 1, 2, ..., M+1)$*,* $\tilde{F_i^j}(1 \le i \le p_j)$ *is a fuzzy set. We call the information system* $S = (U, \tilde{A})$ *a fuzzy decision table.*

Definition 2. *Suppose a fuzzy decision table* $S = (U, \tilde{A})$*.* P, Q *are fuzzy equivalence relations.* $U/IND(P) = \{\tilde{X_1}, \tilde{X_2}, ..., \tilde{X_n}\}$*,* $U/IND(Q) = \{\tilde{Y_1}, \tilde{Y_2}, ..., \tilde{Y_m}\}$*.* $\forall \tilde{X_i} \in U/IND(P)$*,* $\forall \tilde{Y_j} \in U/IND(Q)$ *are all fuzzy sets on* U*, then the entropy of knowledge* P *can be defined as:*

$$H(P) = -\sum_{i=1}^{n} p(\tilde{X_i}) \log p(\tilde{X_i}) = -\sum_{i=1}^{n} \frac{\sum_{k=1}^{|U|} u_{\tilde{X_i}}(x_k)}{|U|} \log \frac{\sum_{k=1}^{|U|} u_{\tilde{X_i}}(x_k)}{|U|} \tag{1}$$

The conditional entropy $H(Q|P)$ *is expressed as:*

$$\begin{aligned} H(Q|P) &= -\sum_{i=1}^{n} p(\tilde{X_i}) \sum_{j=1}^{m} p(\tilde{Y_j}|\tilde{X_i}) \log p(\tilde{Y_j}|\tilde{X_i}) \\ &= -\sum_{i=1}^{n} \frac{\sum_{k=1}^{|U|} u_{\tilde{X_i}(x_k)}}{|U|} \sum_{i=1}^{m} \frac{\sum_{k=1}^{|U|} u_{\tilde{X_i} \cap \tilde{Y_j}}(x_k)}{\sum_{k=1}^{|U|} u_{\tilde{X_i}(x_k)}} \log \frac{\sum_{k=1}^{|U|} u_{\tilde{X_i} \cap \tilde{Y_j}}(x_k)}{\sum_{k=1}^{|U|} u_{\tilde{X_i}(x_k)}}. \end{aligned} \tag{2}$$

$U/IND(P) = \bigotimes U/IND\{\tilde{A^j}\}, \tilde{A^j} \in P$, $U/IND(Q) = \bigotimes U/IND\{\tilde{A^j}\}, \tilde{A^j} \in Q$. And $\tilde{T_1} \bigotimes \tilde{T_2} = \{\tilde{X} \cap \tilde{Y} : \forall \tilde{X} \in \tilde{T_1}, \forall \tilde{Y} \in \tilde{T_2}, \tilde{X} \cap \tilde{Y} \neq \emptyset\}$. Moreover, $u(.)$ is the membership function of a fuzzy set. $u_{\tilde{T_1} \cap \tilde{T_2} \cap ... \cap \tilde{T_n}} = min\{u_{\tilde{T_1}}(x), u_{\tilde{T_2}}(x), ..., u_{\tilde{T_n}}(x)\}$, $\tilde{T_i}$ is the fuzzy set on U.

For all the $\tilde{A^j} \in C - \Re$, the significance $SGF(\tilde{A^j}, \Re, D)$ could be expressed as:

$$SGF(\tilde{A^j}, \Re, D) = I(\Re \cup \{\tilde{A^j}\}; D) - I(\Re; D) = H(D|\Re) - H(D|\Re \cup |\{\tilde{A^j}\}) \quad (3)$$

Attribute $\tilde{A^j}$ is more important as the value of $SGF(\tilde{A^j}, \Re, D)$ increases. The algorithmic procedure of mutual information-based algorithm for fuzzy-rough feature selection(MIFRFS) has been put forward in [6].

3 Proposed Feature Selection Algorithm

As the number of selected features increases, it is costly to compute the conditional mutual information between selected features and the decision feature. An alternative approach is to select features based on maximal relevance criterion[7]. Max-Relevance is to search features satisfying (4), which approximates $I(\Re; D)$ with the mean value of the mutual information between individual feature and target class label.

$$\max R(\Re, D), R = \frac{1}{|\Re|} \sum_{\tilde{A^j} \in \Re} I(\tilde{A^j}; D) \quad (4)$$

It is likely that features selected according to Max-Relevance could have rich redundancy, that is, the dependency among these features could be large. When two features highly depend on each other, the respective class discriminative power would not change much if one of them is removed. Therefore, the following maximal independence condition can be added to select mutually exclusive features:

$$\max S(\Re, D),$$
$$S = \frac{1}{|\Re(\Re-1)|} \sum_{\tilde{A^i} != \tilde{A^j} \in \Re, i<j} (I(\tilde{A^j}; D|\tilde{A^i}) + I(\tilde{A^i}; D|\tilde{A^j})) \quad (5)$$

The operator $\phi(\Re, S)$ is defined to combine R and S, and the following simplest form is considered to optimize R and S simultaneously:

$$\max \phi(\Re, S), \phi = \Re + S. \quad (6)$$

In practice, incremental search methods can be used to find the near-optimal features defined by $\phi(\cdot)$[7]. Given the feature set $\Re_{d-1}$ with $d-1$ features, the task is to select the dth feature from the set $\{C - \Re_{d-1}\}$. This is done by selecting the feature that maximizes $\phi(\cdot)$. The respective incremental algorithm optimizes the following condition:

$$\max_{\tilde{A^i} \in C - \Re_{d-1}} [I(\tilde{A^j}; D) + \frac{1}{d-1} \sum_{\tilde{A^j} \in \Re_{d-1}} I(\tilde{A^j}; D|\tilde{A^i})] \quad (7)$$

To use the criterion above instead of calculating the conditional mutual information between all of the selected features and the decision feature, the proposed algorithm is as follows.

Algorithm: Improved mutual information based fuzzy rough feature selection

Step1. Let $\Re = \emptyset$, for each conditional attributes $C - \Re$ do{

1. For every attribute $\tilde{A}^j \in C - \Re$, and every attribute $\tilde{A}^i \in \Re$, compute the sum of conditional mutual information $\sum_{\tilde{A}^j \in \Re_{d-1}} I(\tilde{A}^j; D|\tilde{A}^i)$;
2. Select the attribute which brings the maximum of $I(\tilde{A}^j; D)+ \frac{1}{d-1} \sum_{\tilde{A}^j \in \Re_{d-1}} I(\tilde{A}^j; D|\tilde{A}^i)$, then record it as $\tilde{A}^j$(if exists multi attributes achieving the maximum at the same time, choose one having the least number of fuzzy classes as $\tilde{A}^j$);
3. if $I(\Re; D) = I(\Re \cup \tilde{A}^j; D)$, break; otherwise, return to 1), $\Re \Leftarrow \Re \cup \{\tilde{A}^j\}$;

}

Step2. Condition attribute set $\Re$ is a relative reduction.

In this regard, it should be noted that the minimum-redundancy-maximum-relevance (mRMR) based feature selection algorithm[7] selects a subset of features from the whole feature set by maximizing the relevance and minimizing the redundancy of the selected features. However, the redundancy measure of the mRMR method does not take into account the supervised information of class labels, while both relevance and independence criteria of the proposed method are computed based on the class labels. Hence, the proposed method in this article provides better performance than the existing mRMR method. In addition, the MRMS proposed in paper[8] lacks of the intuitive interpretation for the computation under an algebraic setting. This paper investigates a fast feature selection method based on mutual information with fuzzy rough approach.

4 Experiments and Results

The performance of the proposed method is extensively studied and compared with MIFRFS algorithm in this section. Both of the algorithms are implemented in matlab language and run in Windows environment having machine configuration Intel Core2 T5500, 1.66 GHz, 2 MB cache, and 1 GB RAM. To analyze the performance of different algorithms, the experimentation is done on the dataset from Wujing Power Plant, which contains 6905 objects and 174 condition attributes, 1 decision attribute(the consumption of coal). The dataset is chosen under consideration of steady state. The major metrics for evaluating the

performance of different algorithms are the classification accuracy of 1-nearest neighbor (1-NN) rule. To compute the prediction accuracy of 1-NN rule, several training and testing sets are performed on the dataset.

The experiments evaluate the performance of the two algorithms with different number of fuzzy intervals (including the number of fuzzy classes of condition attributes and classes of decision attributes). As an example, Fig. 1(a) demonstrates the efficiency of the proposed method, compared to MIFRFS approach. The elapsed time changes when the number of classes of decision attribute increases. The experiments are conducted on the condition with all the condition attributes are fuzzified to 2 classes.

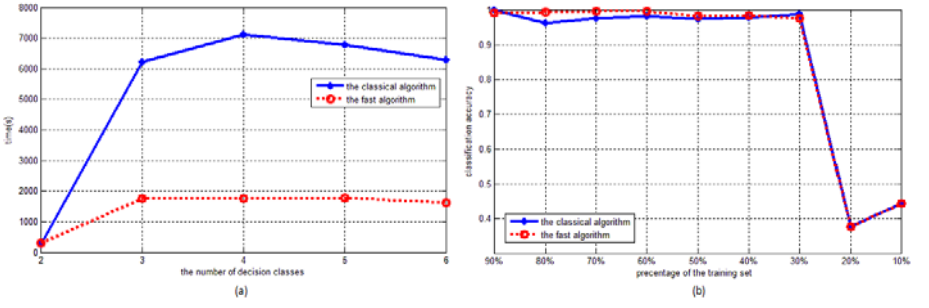

Fig. 1. Elapsed time and classification accuracy on coal-fired power units

Obviously, our proposed method has a much better performance on efficiency. In the context of discriminant analysis, whether the selected feature subset can provide good generalization ability to the classifier or not is very important.In our study, we use 1NN to assess the performance of different feature selection algorithms. Different proportions of training samples are used to train and the samples not contained in the training set construct the corresponding testing set (Fig. 1(b)). Results suggest that the classification accuracies of both of the two methods decrease as the number of classes of the decision attribute increases to a trend, but it is independent to the number of fuzzy classes of condition attributes. It can be easily inferred from theory. As an example, the experiment is performed on the condition that condition attributes are partitioned to 5 classes and the decision attribute is partitioned to 2 classes. All the training sets are chosen randomly, and the vertical axis represents the mean accuracy with the same proportion by 10 times.

The classification achieves higher accuracy when the decision attribute is partitioned to 2 classes. Only 2 or 3 features are extracted. From Fig. 1(b), the classification accuracy of both of the two methods falls sharply when the samples of training set are below 20% of the dataset. Experimental results show that both of the two methods have similar classification accuracies and not far selected features.

5 Conclusions and Future Work

The approach to fuzzy rough feature selection has a costly computation so that it can not be worked out when the number of selected features is large. In this paper, an approximate expression is proposed to decrease the complexity. The corresponding algorithm is also presented. Experiments demonstrate that both of the two methods can preserve the ability of classification and achieve a high accuracy when partitioned to 2 classes. In addition, the performance test shows that our proposed method takes much less time, obtaining expected effectiveness of fuzzy rough approach.

As the number of classes of the decision attribute increases, the number of selected features is also increasing. Further experiments will be done by the increasing number of selected features under more classes of the decision attribute. In this case, the selected features can be used for energy-saving and consumption-lowering for the units.

Acknowledgment. This work is supported by Sub-project Research (2009CB219801) of National Key Basic Research Program and Natural Science Foundation of China (No.60863001, No.61073189) and Innovation Program of Shanghai Municipal Education Commission (No.10ZZ115) and a grant from Shanghai University of Electric Power (A-3101-11-002) and supported by the Research Award Fund for Outstanding Young Teachers in Higher Education Institutions, Shanghai (B211058K).

References

1. Yu, L., Liu, H.: Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In: Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC (2003)
2. Jensen, R., Shen, Q.: New Approaches to Fuzzy-Rough Feature Selection. IEEE Transactions on Fuzzy Systems 17(4), 824–838 (2009)
3. Hu, Q.H., Xie, Z.X., Yu, D.R.: Hybrid Attribute Reduction Based on A Novel Fuzzy-rough Model and Information Granulation. Pattern Recognition 40, 3509–3521 (2007)
4. Jensen, R., Shen, Q.: Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches. Wiley-IEEE Press (2008)
5. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer, Norwell (1991)
6. Xu, F.F., Miao, D.Q., Wei, L.: Fuzzy-rough Attribute Reduction via Mutual Information with An Application to Cancer Classification. Computer and Mathematics with Applications 57, 1010–1017 (2009)
7. Peng, H., Long, F., Ding, C.: Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8), 1226–1238 (2005)
8. Maji, P., Paul, S.: Rough Set Based Maximum Relevance-Maximum Significance Criterion and Gene Selection from Microarray Data. Int. J. Approx. Reason. (2010) (in press)